CAPÍTULO 1

Introducción a la estadística y al análisis de datos

1.1 Panorama general: inferencia estadística, muestras, poblaciones y el papel de la probabilidad

Desde inicios de la década de los ochenta del siglo pasado y hasta lo que ha transcurrido del siglo XXI la industria estadounidense ha puesto una enorme atención en el *mejoramiento de la calidad*. Se ha dicho y escrito mucho acerca del "milagro industrial" en Japón, que comenzó a mediados del siglo XX. Los japoneses lograron el éxito en donde otras naciones fallaron, a saber, en la creación de un entorno que permita la manufactura de productos de alta calidad. Gran parte del éxito de los japoneses se atribuye al uso de *métodos estadísticos* y del pensamiento estadístico entre el personal gerencial.

Empleo de datos científicos

El uso de métodos estadísticos en la manufactura, el desarrollo de productos alimenticios, el software para computadoras, las fuentes de energía, los productos farmacéuticos y muchas otras áreas implican el acopio de información o **datos científicos**. Por supuesto que la obtención de datos no es algo nuevo, ya que se ha realizado por más de mil años. Los datos se han recabado, resumido, reportado y almacenado para su examen cuidadoso. Sin embargo, hay una diferencia profunda entre el acopio de información científica y la **estadística inferencial**. Esta última ha recibido atención legítima en décadas recientes.

La estadística inferencial generó un número enorme de "herramientas" de los métodos estadísticos que utilizan los profesionales de la estadística. Los métodos estadísticos se diseñan para contribuir al proceso de realizar juicios científicos frente a la **incertidumbre** y a la **variación**. Dentro del proceso de manufactura, la densidad de producto de un material específico no siempre será la misma. De hecho, si un proceso es discontinuo en vez de continuo, la densidad de material no sólo variará entre los lotes que salen de la línea de producción (variación de un lote a otro), sino también dentro de los propios lotes. Los métodos estadísticos se utilizan para analizar datos de procesos como el anterior; el objetivo de esto es tener una mejor orientación respecto de cuáles cambios se deben realizar en el proceso para mejorar su **calidad**. En este proceso la calidad bien podría

www.FreeLibros.me

definirse en relación con su grado de acercamiento a un valor de densidad meta en armonía con *qué parte de las veces* se cumple este criterio de cercanía. A un ingeniero podría interesarle un instrumento específico que se utilice para medir el monóxido de azufre en estudios sobre la contaminación atmosférica. Si el ingeniero dudara respecto de la eficacia del instrumento, tendría que tomar en cuenta dos **fuentes de variación**. La primera es la variación en los valores del monóxido de azufre que se encuentran en el mismo lugar el mismo día. La segunda es la variación entre los valores observados y la cantidad **real** de monóxido de azufre que haya en el aire en ese momento. Si cualquiera de estas dos fuentes de variación es excesivamente grande (según algún estándar determinado por el ingeniero), quizá se necesite remplazar el instrumento. En un estudio biomédico de un nuevo fármaco que reduce la hipertensión, 85% de los pacientes experimentaron alivio; aunque por lo general se reconoce que el medicamento actual o el "viejo" alivia a 80% de los pacientes que sufren hipertensión crónica. Sin embargo, el nuevo fármaco es más caro de elaborar y podría tener algunos efectos colaterales. ¿Se debería adoptar el nuevo medicamento? Éste es un problema con el que las empresas farmacéuticas, junto con la FDA (Federal Drug Administration), se encuentran a menudo (a veces es mucho más complejo). De nuevo se debe tomar en cuenta las necesidades de variación. El valor del "85%" se basa en cierto número de pacientes seleccionados para el estudio. Tal vez si se repitiera el estudio con nuevos pacientes ¡el número observado de "éxitos" sería de 75%! Se trata de una variación natural de un estudio a otro que se debe tomar en cuenta en el proceso de toma de decisiones. Es evidente que tal variación es importante, ya que la variación de un paciente a otro es endémica al problema.

Variabilidad en los datos científicos

En los problemas analizados anteriormente los métodos estadísticos empleados tienen que ver con la variabilidad y en cada caso la variabilidad que se estudia se encuentra en datos científicos. Si la densidad del producto observada en el proceso fuera siempre la misma y siempre fuera la esperada, no habría necesidad de métodos estadísticos. Si el dispositivo para medir el monóxido de azufre siempre diera el mismo valor y éste fuera exacto (es decir, correcto), no se requeriría análisis estadístico. Si entre un paciente y otro no hubiera variabilidad inherente a la respuesta al medicamento (es decir, si el fármaco siempre causara alivio o nunca aliviara), la vida sería muy sencilla para los científicos de las empresas farmacéuticas y de la FDA, y los estadísticos no serían necesarios en el proceso de toma de decisiones. Los investigadores de la estadística han originado un gran número de métodos analíticos que permiten efectuar análisis de datos obtenidos de sistemas como los descritos anteriormente, lo cual refleja la verdadera naturaleza de la ciencia que conocemos como estadística inferencial, a saber, el uso de técnicas que, al permitirnos obtener conclusiones (o inferencias) sobre el sistema científico, nos permiten ir más allá de sólo reportar datos. Los profesionales de la estadística usan leyes fundamentales de probabilidad e inferencia estadística para sacar conclusiones respecto de los sistemas científicos. La información se colecta en forma de **muestras** o conjuntos de **observaciones**. En el capítulo 2 se introduce el proceso de muestreo, el cual se continúa analizando a lo largo de todo el libro.

Las muestras se reúnen a partir de **poblaciones**, que son conjuntos de todos los individuos o elementos individuales de un tipo específico. A veces una población representa un sistema científico. Por ejemplo, un fabricante de tarjetas para computadora podría desear eliminar defectos. Un proceso de muestreo implicaría recolectar información de 50 tarjetas de computadora tomadas aleatoriamente durante el proceso. En este caso la población

www.FreeLibros.me

sería representada por todas las tarjetas de computadora producidas por la empresa en un periodo específico. Si se lograra mejorar el proceso de producción de las tarjetas para computadora y se reuniera una segunda muestra de tarjetas, cualquier conclusión que se obtuviera respecto de la efectividad del cambio en el proceso debería extenderse a toda la población de tarjetas para computadora que se produzcan en el "proceso mejorado". En un experimento con fármacos se toma una muestra de pacientes y a cada uno se le administra un medicamento específico para reducir la presión sanguínea. El interés se enfoca en obtener conclusiones sobre la población de quienes sufren hipertensión. A menudo, cuando la planeación ocupa un lugar importante en la agenda, es muy importante el acopio de datos científicos en forma sistemática. En ocasiones la planeación está, por necesidad, bastante limitada. Con frecuencia nos enfocamos en ciertas propiedades o características de los elementos u objetos de la población. Cada característica tiene importancia de ingeniería específica o, digamos, biológica para el "cliente", el científico o el ingeniero que busca aprender algo acerca de la población. Por ejemplo, en uno de los casos anteriores la calidad del proceso se relacionaba con la densidad del producto al salir del proceso. Un(a) ingeniero(a) podría necesitar estudiar el efecto de las condiciones del proceso, la temperatura, la humedad, la cantidad de un ingrediente particular, etcétera. Con ese fin podría mover de manera sistemática estos **factores** a cualesquiera niveles que se sugieran, de acuerdo con cualquier prescripción o **diseño experimental** que se desee. Sin embargo, un científico silvicultor que está interesado en estudiar los factores que influyen en la densidad de la madera en cierta clase de árbol no necesariamente tiene que diseñar un experimento. Este caso quizá requiera un **estudio observacional**, en el cual los datos se acopian en el campo pero no es posible seleccionar de antemano los **niveles de los factores**. Ambos tipos de estudio se prestan a los métodos de la inferencia estadística. En el primero, la calidad de las inferencias dependerá de la planeación adecuada del experimento. En el segundo, el científico está a expensas de lo que pueda recopilar. Por ejemplo, si un agrónomo se interesara en estudiar el efecto de la lluvia sobre la producción de plantas sería lamentable que recopilara los datos durante una sequía.

Es bien conocida la importancia del pensamiento estadístico para los administradores y el uso de la inferencia estadística para el personal científico. Los investigadores obtienen mucho de los datos científicos. Los datos proveen conocimiento acerca del fenómeno científico. Los ingenieros de producto y de procesos aprenden más en sus esfuerzos fuera de línea para mejorar el proceso. También logran una comprensión valiosa al reunir datos de producción (supervisión en línea) sobre una base regular, lo cual les permite determinar las modificaciones que se requiere realizar para mantener el proceso en el nivel de calidad deseado.

En ocasiones un científico sólo desea obtener alguna clase de resumen de un conjunto de datos representados en la muestra. En otras palabras, no requiere estadística inferencial. En cambio, le sería útil un conjunto de estadísticos o la **estadística descriptiva**. Tales números ofrecen un sentido de la ubicación del centro de los datos, de la variabilidad en los datos y de la naturaleza general de la distribución de observaciones en la muestra. Aunque no se incorporen métodos estadísticos específicos que lleven a la **inferencia estadística**, se puede aprender mucho. A veces la estadística descriptiva va acompañada de gráficas. El software estadístico moderno permite el cálculo de **medias**, **medianas**, **desviaciones estándar** y otros estadísticos de una sola cifra, así como el desarrollo de gráficas que presenten una "huella digital" de la naturaleza de la muestra. En las secciones siguientes veremos definiciones e ilustraciones de los estadísticos y descripciones de recursos gráficos como histogramas, diagramas de tallo y hojas, diagramas de dispersión, gráficas de puntos y diagramas de caja.

www.FreeLibros.me

El papel de la probabilidad

En los capítulos 2 a 6 de este libro se presentan los conceptos fundamentales de la probabilidad. Un estudio concienzudo de las bases de tales conceptos permitirá al lector comprender mejor la inferencia estadística. Sin algo de formalismo en teoría de probabilidad, el estudiante no podría apreciar la verdadera interpretación del análisis de datos a través de los métodos estadísticos modernos. Es muy natural estudiar probabilidad antes de estudiar inferencia estadística. Los elementos de probabilidad nos permiten cuantificar la fortaleza o "confianza" en nuestras conclusiones. En este sentido, los conceptos de probabilidad forman un componente significativo que complementa los métodos estadísticos y ayuda a evaluar la consistencia de la inferencia estadística. Por consiguiente, la disciplina de la probabilidad brinda la transición entre la estadística descriptiva y los métodos inferenciales. Los elementos de la probabilidad permiten expresar la conclusión en el lenguaje que requieren los científicos y los ingenieros. El ejemplo que sigue permite al lector comprender la noción de un valor-*P*, el cual a menudo proporciona el "fundamento" en la interpretación de los resultados a partir del uso de métodos estadísticos.

Ejemplo 1.1: Suponga que un ingeniero se encuentra con datos de un proceso de producción en el cual se muestrean 100 artículos y se obtienen 10 defectuosos. Se espera y se anticipa que ocasionalmente habrá artículos defectuosos. Obviamente estos 100 artículos representan la muestra. Sin embargo, se determina que, a largo plazo, la empresa sólo puede tolerar 5% de artículos defectuosos en el proceso. Ahora bien, los elementos de probabilidad permiten al ingeniero determinar qué tan concluyente es la información muestral respecto de la naturaleza del proceso. En este caso la **población** representa conceptualmente todos los artículos posibles en el proceso. Suponga que averiguamos que, *si el proceso es aceptable,* es decir, que su producción no excede un 5% de artículos defectuosos, hay una probabilidad de 0.0282 de obtener 10 o más artículos defectuosos en una muestra aleatoria de 100 artículos del proceso. Esta pequeña probabilidad sugiere que, en realidad, a largo plazo el proceso tiene un porcentaje de artículos defectuosos mayor al 5%. En otras palabras, en las condiciones de un proceso aceptable casi nunca se obtendría la información muestral que se obtuvo. Sin embargo, ¡se obtuvo! Por lo tanto, es evidente que la probabilidad de que se obtuviera sería mucho mayor si la tasa de artículos defectuosos del proceso fuera mucho mayor que 5%. ▟

A partir de este ejemplo se vuelve evidente que los elementos de probabilidad ayudan a traducir la información muestral en algo concluyente o no concluyente acerca del sistema científico. De hecho, lo aprendido probablemente constituya información inquietante para el ingeniero o administrador. Los métodos estadísticos (que examinaremos con más detalle en el capítulo 10) produjeron un valor-*P* de 0.0282. El resultado sugiere que es **muy probable que el proceso no sea aceptable**. En los capítulos siguientes se trata detenidamente el concepto de **valor-*P***. El próximo ejemplo brinda una segunda ilustración.

Ejemplo 1.2: Con frecuencia, la naturaleza del estudio científico señalará el papel que desempeñan la probabilidad y el razonamiento deductivo en la inferencia estadística. El ejercicio 9.40 en la página 294 proporciona datos asociados con un estudio que se llevó a cabo en el Virginia Polytechnic Institute and State University acerca del desarrollo de una relación entre las raíces de los árboles y la acción de un hongo. Los minerales de los hongos se transfieren a los árboles, y los azúcares de los árboles a los hongos. Se plantaron dos muestras de 10 plantones de roble rojo norteño en un invernadero, una de ellas contenía

www.FreeLibros.me

plantones tratados con nitrógeno y la otra plantones sin tratamiento. Todas las demás condiciones ambientales se mantuvieron constantes. Todos los plantones contenían el hongo *Pisolithus tinctorus.* En el capítulo 9 se incluyen más detalles. Los pesos en gramos de los tallos se registraron después de 140 días y los datos se presentan en la tabla 1.1.

Tabla 1.1: Conjunto de datos del ejemplo 1.2

Sin nitrógeno	**Con nitrógeno**
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

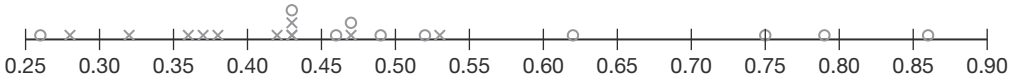


Figura 1.1: Gráfica de puntos de los datos de peso del tallo.

En este ejemplo hay dos muestras tomadas de dos **poblaciones distintas**. El objetivo del experimento es determinar si el uso del nitrógeno influye en el crecimiento de las raíces. Éste es un estudio comparativo (es decir, es un estudio en el que se busca comparar las dos poblaciones en cuanto a ciertas características importantes). Los datos se deben graficar como se indica en el diagrama de puntos de la figura 1.1. Los valores $\circ$ representan los datos "con nitrógeno" y los valores $\times$ los datos "sin nitrógeno".

Observe que la apariencia general de los datos podría sugerir al lector que, en promedio, el uso del nitrógeno aumenta el peso del tallo. Cuatro observaciones con nitrógeno son considerablemente más grandes que cualquiera de las observaciones sin nitrógeno. La mayoría de las observaciones sin nitrógeno parece estar por debajo del centro de los datos. La apariencia del conjunto de datos parece indicar que el nitrógeno es efectivo. Pero, ¿cómo se cuantifica esto? ¿Cómo se puede resumir toda la evidencia visual aparente de manera que tenga algún significado? Como en el ejemplo anterior, se pueden utilizar los fundamentos de la probabilidad. Las conclusiones se resumen en una declaración de probabilidad o valor-*P.* Aquí no demostraremos la inferencia estadística que produce la probabilidad resumida. Igual que en el ejemplo 1.1, tales métodos se estudiarán en el capítulo 10. El problema gira alrededor de la "probabilidad de que datos como éstos se puedan observar", *dado que el nitrógeno no tiene efecto*; en otras palabras, dado que ambas muestras se generaron a partir de la misma población. Suponga que esta probabilidad es pequeña, digamos de 0.03; un porcentaje que podría constituir suficiente evidencia de que el uso del nitrógeno en realidad influye en el peso promedio del tallo en los plantones de roble rojo (aparentemente lo aumenta). ▟

www.FreeLibros.me

¿Cómo trabajan juntas la probabilidad y la inferencia estadística?

Es importante para el lector que comprenda claramente la diferencia entre la disciplina de la probabilidad, una ciencia por derecho propio, y la disciplina de la estadística inferencial. Como señalamos, el uso o la aplicación de conceptos de probabilidad permite interpretar la vida cotidiana a partir de los resultados de la inferencia estadística. En consecuencia, se afirma que la inferencia estadística emplea los conceptos de probabilidad. A partir de los dos ejemplos anteriores aprendimos que la información muestral está disponible para el analista y que, con la ayuda de métodos estadísticos y elementos de probabilidad, podemos obtener conclusiones acerca de alguna característica de la población (en el ejemplo 1.1 el proceso al parecer no es aceptable, y en el ejemplo 1.2 parece ser que el nitrógeno en verdad influye en el peso promedio de los tallos). Así, para un problema estadístico, **la muestra, junto con la estadística inferencial, nos permite obtener conclusiones acerca de la población, ya que la estadística inferencial utiliza ampliamente los elementos de probabilidad**. Tal razonamiento es *inductivo* por naturaleza. Ahora, cuando avancemos al capítulo 2 y los siguientes, el lector encontrará que, a diferencia de lo que hicimos en nuestros dos ejemplos actuales, no nos enfocaremos en resolver problemas estadísticos. En muchos de los ejemplos que estudiaremos no utilizaremos muestras. Lo que haremos será describir claramente una población con todas sus características conocidas. Las preguntas importantes se enfocarán en la naturaleza de los datos que hipotéticamente se podrían obtener a partir de la población. Entonces, podríamos afirmar que los **elementos de probabilidad nos permiten sacar conclusiones acerca de las características de los datos hipotéticos que se tomen de la población, con base en las características conocidas de la población**. Esta clase de razonamiento es *deductivo* por naturaleza. La figura 1.2 muestra la relación básica entre la probabilidad y la estadística inferencial.

Figura 1.2: Relación básica entre la probabilidad y la estadística inferencial.

Ahora bien, en términos generales, ¿cuál campo es más importante, el de la probabilidad o el de la estadística? Ambos son muy importantes y evidentemente se complementan. La única certeza respecto de la didáctica de ambas disciplinas radica en el hecho de que, si la estadística se debe enseñar con un nivel mayor al de un simple "libro de cocina", entonces hay que comenzar por enseñar la disciplina de la probabilidad. Esta regla se basa en el hecho de que un analista no podrá aprender nada sobre una población a partir de una muestra hasta que aprenda los rudimentos de incertidumbre en esa muestra. Considere el ejemplo 1.1; en el que la pregunta se centra en si la población, definida por el proceso, tiene o no más de 5% de elementos defectuosos. En otras palabras, la suposición es que 5 de cada 100 artículos, **en promedio,** salen defectuosos. Ahora bien, la muestra contiene 100 artículos y 10 están defectuosos. ¿Esto apoya o refuta la supo-

www.FreeLibros.me

sición? Aparentemente la refuta porque 10 artículos de cada 100 parecen ser "un trozo grande". ¿Pero cómo podríamos saber esto sin tener nociones de probabilidad? La única manera en que podremos aprender las condiciones en las cuales el proceso es aceptable (5% de defectuosos) es estudiando el material de los siguientes capítulos. La probabilidad de obtener 10 o más artículos defectuosos en una muestra de 100 es de 0.0282.

Dimos dos ejemplos en donde los elementos de probabilidad ofrecen un resumen que el científico o el ingeniero pueden usar como evidencia para basar una decisión. El puente entre los datos y la conclusión está, por supuesto, basado en los fundamentos de la inferencia estadística, la teoría de la distribución y las distribuciones de muestreos que se examinarán en capítulos posteriores.

1.2 Procedimientos de muestreo; recolección de los datos

En la sección 1.1 estudiamos muy brevemente el concepto de muestreo y el proceso de muestreo. Aunque el muestreo parece ser un concepto simple, la complejidad de las preguntas que se deben contestar acerca de la población, o las poblaciones, en ocasiones requiere que el proceso de muestreo sea muy complejo. El concepto de muestreo se examinará de manera técnica en el capítulo 8, pero aquí nos esforzaremos por dar algunas nociones de sentido común sobre el muestreo. Ésta es una transición natural hacia el análisis del concepto de variabilidad.

Muestreo aleatorio simple

La importancia del muestreo adecuado gira en torno al grado de confianza con que el analista es capaz de responder las preguntas que se plantean. Supongamos que sólo hay una población en el problema. Recuerde que en el ejemplo 1.2 había dos poblaciones implicadas. El **muestreo aleatorio simple** significa que cierta muestra dada de un *tamaño muestral* específico tiene la misma probabilidad de ser seleccionada que cualquiera otra muestra del mismo tamaño. El término **tamaño muestral** simplemente indica el número de elementos en la muestra. Evidentemente, en muchos casos se puede utilizar una tabla de números aleatorios para seleccionar la muestra. La ventaja del muestreo aleatorio simple radica en que ayuda a eliminar el problema de tener una muestra que refleje una población diferente (quizá más restringida) de aquella sobre la cual se necesitan realizar las inferencias. Por ejemplo, se elige una muestra para contestar diferentes preguntas respecto de las preferencias políticas en cierta entidad de Estados Unidos. La muestra implica la elección de, digamos, 1 000 familias y una encuesta a aplicar. Ahora bien, suponga que no se utiliza el muestreo aleatorio, sino que todas o casi todas las 1 000 familias se eligen de una zona urbana. Se considera que las preferencias políticas en las áreas rurales difieren de las de las áreas urbanas. En otras palabras, la muestra obtenida en realidad confinó a la población y, por lo tanto, las inferencias también se tendrán que restringir a la "población confinada", y en este caso el confinamiento podría resultar indeseable. Si, de hecho, se necesitara hacer las inferencias respecto de la entidad como un todo, a menudo se diría que la muestra con un tamaño de 1 000 familias aquí descrita es una **muestra sesgada**.

Como antes sugerimos, el muestreo aleatorio simple no siempre es adecuado. El enfoque alternativo que se utilice dependerá de la complejidad del problema. Con frecuencia, por ejemplo, las unidades muestrales no son homogéneas y se dividen naturalmente en grupos que no se traslapan y que son homogéneos. Tales grupos se llaman *estratos*, y

www.FreeLibros.me

un procedimiento llamado *muestreo aleatorio estratificado* implica la selección al azar de una muestra *dentro* de cada estrato. El propósito de esto es asegurarse de que ninguno de los estratos esté sobrerrepresentado ni subrepresentado. Por ejemplo, suponga que se aplica una encuesta a una muestra para reunir opiniones preliminares respecto de un referéndum que se piensa realizar en determinada ciudad. La ciudad está subdividida en varios grupos étnicos que representan estratos naturales y, para no excluir ni sobrerrepresentar a algún grupo de cada uno de ellos, se eligen muestras aleatorias separadas de cada grupo.

Diseño experimental

El concepto de aleatoriedad o asignación aleatoria desempeña un papel muy importante en el área del **diseño experimental**, que se presentó brevemente en la sección 1.1 y es un fundamento muy importante en casi cualquier área de la ingeniería y de la ciencia experimental. Estudiaremos este tema con detenimiento en los capítulos 13 a 15. Sin embargo, es conveniente introducirlo aquí brevemente en el contexto del muestreo aleatorio. Un conjunto de los llamados **tratamientos** o **combinaciones de tratamientos** se vuelven las poblaciones que se van a estudiar o a comparar en algún sentido. Un ejemplo es el tratamiento "con nitrógeno" *versus* "sin nitrógeno" del ejemplo 1.2. Otro ejemplo sencillo sería "placebo" *versus* "medicamento activo" o, en un estudio sobre la fatiga por corrosión, tendríamos combinaciones de tratamientos que impliquen especímenes con recubrimiento o sin recubrimiento, así como condiciones de alta o de baja humedad, a las cuales se somete el espécimen. De hecho, habrían cuatro combinaciones de factores o de tratamientos (es decir, 4 poblaciones), y se podrían formular y responder muchas preguntas científicas usando los métodos estadísticos e inferenciales. Considere primero la situación del ejemplo 1.2. En el experimento hay 20 plantones enfermos implicados. A partir de los datos es fácil observar que los plantones son diferentes entre sí. Dentro del grupo tratado con nitrógeno (o del grupo que no se trató con nitrógeno) hay **variabilidad** considerable en el peso de los tallos, la cual se debe a lo que por lo general se denomina **unidad experimental**. Éste es un concepto tan importante en la estadística inferencial que no es posible describirlo totalmente en este capítulo. La naturaleza de la variabilidad es muy importante. Si es demasiado grande, debido a que resulta de una condición de excesiva falta de homogeneidad en las unidades experimentales, la variabilidad "eliminará" cualquier diferencia detectable entre ambas poblaciones. Recuerde que en este caso eso no ocurrió.

La gráfica de puntos de la figura 1.1 y el valor-P indican una clara distinción entre esas dos condiciones. ¿Qué papel desempeñan tales unidades experimentales en el proceso mismo de recolección de los datos? El enfoque por sentido común y, de hecho, estándar, es asignar los 20 plantones o unidades experimentales **aleatoriamente a las dos condiciones o tratamientos**. En el estudio del medicamento podríamos decidir utilizar un total de 200 pacientes disponibles, quienes serán claramente distinguibles en algún sentido. Ellos son las unidades experimentales. No obstante, tal vez todos tengan una condición crónica que podría ser tratada con el fármaco. Así, en el denominado **diseño completamente aleatorio**, se asignan al azar 100 pacientes al placebo y 100 al medicamento activo. De nuevo, son estas unidades experimentales en el grupo o tratamiento las que producen la variabilidad en el resultado de los datos (es decir, la variabilidad en el resultado medido), digamos, de la presión sanguínea o cualquier valor de la eficacia de un medicamento que sea importante. En el estudio de la fatiga por corrosión las unidades experimentales son los especímenes que se someten a la corrosión.

www.FreeLibros.me

¿Por qué las unidades experimentales se asignan aleatoriamente?

¿Cuál es el posible efecto negativo de no asignar aleatoriamente las unidades experimentales a los tratamientos o a las combinaciones de tratamientos? Esto se observa más claramente en el caso del estudio del medicamento. Entre las características de los pacientes que producen variabilidad en los resultados están la edad, el género y el peso. Tan sólo suponga que por casualidad el grupo del placebo contiene una muestra de personas que son predominantemente más obesas que las del grupo del tratamiento. Quizá los individuos más obesos muestren una tendencia a tener una presión sanguínea más elevada, lo cual evidentemente sesgará el resultado y, por lo tanto, cualquier resultado que se obtenga al aplicar la inferencia estadística podría tener poco que ver con el efecto del medicamento, pero mucho con las diferencias en el peso de ambas muestras de pacientes.

Deberíamos enfatizar la importancia del término **variabilidad**. La variabilidad excesiva entre las unidades experimentales "disfraza" los hallazgos científicos. En secciones posteriores intentaremos clasificar y cuantificar las medidas de variabilidad. En las siguientes secciones presentaremos y analizaremos cantidades específicas que se calculan en las muestras; las cantidades proporcionan una idea de la naturaleza de la muestra respecto de la ubicación del centro de los datos y la variabilidad de los mismos. Un análisis de varias de tales medidas de un solo número permite ofrecer un preámbulo de que la información estadística será un componente importante de los métodos estadísticos que se utilizarán en capítulos posteriores. Estas medidas, que ayudan a clasificar la naturaleza del conjunto de datos, caen en la categoría de **estadísticas descriptivas**. Este material es una introducción a una presentación breve de los métodos pictóricos y gráficos que van incluso más allá en la caracterización del conjunto de datos. El lector debería entender que los métodos estadísticos que se presentan aquí se utilizarán a lo largo de todo el texto. Para ofrecer una imagen más clara de lo que implican los estudios de diseño experimental se presenta el ejemplo 1.3.

Ejemplo 1.3: Se realizó un estudio sobre la corrosión con la finalidad de determinar si al recubrir una aleación de aluminio con una sustancia retardadora de la corrosión, el metal se corroe menos. El recubrimiento es un protector que los anunciantes afirman que minimiza el daño por fatiga en esta clase de material. La influencia de la humedad sobre la magnitud de la corrosión también es de interés. Una medición de la corrosión puede expresarse en millares de ciclos hasta la ruptura del metal. Se utilizaron dos niveles de recubrimiento: sin recubrimiento y con recubrimiento químico contra la corrosión. También se consideraron dos niveles de humedad relativa, de 20% y 80%, respectivamente.

El experimento implica las cuatro combinaciones de tratamientos que se listan en la siguiente tabla. Se usan ocho unidades experimentales, que son especímenes de aluminio preparados, dos de los cuales se asignan aleatoriamente a cada una de las cuatro combinaciones de tratamiento. Los datos se presentan en la tabla 1.2.

Los datos de la corrosión son promedios de los dos especímenes. En la figura 1.3 se presenta una gráfica con los promedios. Un valor relativamente grande de ciclos hasta la ruptura representa una cantidad pequeña de corrosión. Como se podría esperar, al parecer un incremento en la humedad hace que empeore la corrosión. El uso del procedimiento de recubrimiento químico contra la corrosión parece reducir la corrosión. ▟

En este ejemplo de diseño experimental el ingeniero eligió sistemáticamente las cuatro combinaciones de tratamiento. Para vincular esta situación con los conceptos con los que el lector se ha familiarizado hasta aquí, deberíamos suponer que las condiciones

www.FreeLibros.me

Tabla 1.2: Datos para el ejemplo 1.3

Recubrimiento	Humedad	Promedio de corrosión en miles de ciclos hasta la ruptura
Sin recubrimiento	20%	975
	80%	350
Con recubrimiento	20%	1750
químico contra la corrosión	80%	1550

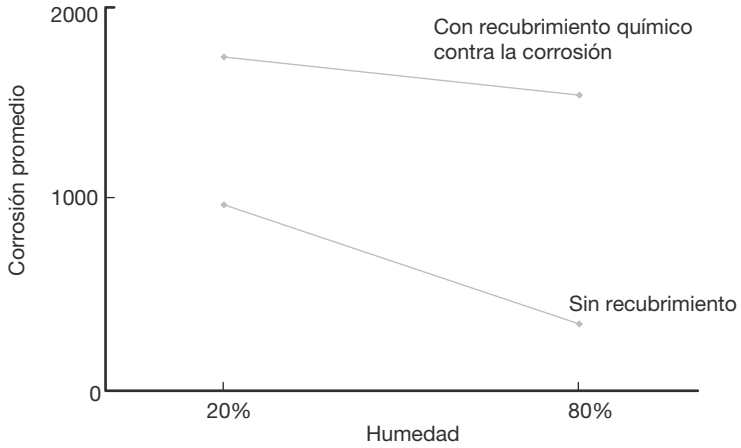


Figura 1.3: Resultados de corrosión para el ejemplo 1.3.

que representan las cuatro combinaciones de tratamientos son cuatro poblaciones separadas y que los dos valores de corrosión observados en cada una de las poblaciones constituyen importantes piezas de información. La importancia del promedio al captar y resumir ciertas características en la población se destacará en la sección 1.3. Aunque a partir de la figura podríamos sacar conclusiones acerca del papel que desempeña la humedad y del efecto de recubrir los especímenes, no podemos evaluar con exactitud los resultados de un punto de vista analítico sin tomar en cuenta la *variabilidad alrededor* del promedio. De nuevo, como señalamos con anterioridad, si los dos valores de corrosión en cada una de las combinaciones de tratamientos son muy cercanos, la imagen de la figura 1.3 podría ser una descripción precisa. Pero si cada valor de la corrosión en la figura es un promedio de dos valores que están ampliamente dispersos, entonces esta variabilidad podría, de hecho, en verdad "eliminar" cualquier información que parezca difundirse cuando tan sólo se observan los promedios. Los siguientes ejemplos ilustran estos conceptos:

1. La asignación aleatoria a las combinaciones de tratamientos (recubrimiento/humedad) de las unidades experimentales (especímenes).
2. El uso de promedios muestrales (valores de corrosión promedio) para resumir la información muestral.
3. La necesidad de considerar las medidas de variabilidad en el análisis de cualquier muestra o conjunto de muestras.

www.FreeLibros.me

Este ejemplo sugiere la necesidad de estudiar el tema que se expone en las secciones 1.3 y 1.4, es decir, el de las estadísticas descriptivas que indican las medidas de la ubicación del centro en un conjunto de datos, y aquellas con las que se mide la variabilidad.

1.3 Medidas de localización: la media y la mediana de una muestra

Las medidas de localización están diseñadas para brindar al analista algunos valores cuantitativos de la ubicación central o de otro tipo de los datos en una muestra. En el ejemplo 1.2 parece que el centro de la muestra con nitrógeno claramente excede al de la muestra sin nitrógeno. Una medida obvia y muy útil es la **media de la muestra**. La media es simplemente un promedio numérico.

Definición 1.1: Suponga que las observaciones en una muestra son $x_1, x_2, \ldots, x_n$. La **media de la muestra**, que se denota con $\bar{x}$, es

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Hay otras medidas de tendencia central que se explican con detalle en capítulos posteriores. Una medida importante es la **mediana de la muestra**. El propósito de la mediana de la muestra es reflejar la tendencia central de la muestra de manera que no sea influida por los valores extremos.

Definición 1.2: Dado que las observaciones en una muestra son $x_1, x_2, \ldots, x_n$, acomodadas en **orden de magnitud creciente**, la mediana de la muestra es

$$\tilde{x} = \begin{cases} x_{(n+1)/2,} & \text{si } n \text{ es impar,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{si } n \text{ es par.} \end{cases}$$

Por ejemplo, suponga que el conjunto de datos es el siguiente: 1.7, 2.2, 3.9, 3.11 y 14.7. La media y la mediana de la muestra son, respectivamente,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

Es evidente que la media es influida de manera considerable por la presencia de la observación extrema, 14.7; en tanto que el lugar de la mediana hace énfasis en el verdadero "centro" del conjunto de datos. En el caso del conjunto de datos de dos muestras del ejemplo 1.2, las dos medidas de tendencia central para las muestras individuales son

$$\begin{aligned} \bar{x} \text{ (sin nitrógeno)} &= 0.399 \text{ gramos}, \\ \tilde{x} \text{ (sin nitrógeno)} &= \frac{0.38 + 0.42}{2} = 0.400 \text{ gramos}, \\ \bar{x} \text{ (con nitrógeno)} &= 0.565 \text{ gramos}, \\ \tilde{x} \text{ (con nitrógeno)} &= \frac{0.49 + 0.52}{2} = 0.505 \text{ gramos}. \end{aligned}$$

Es evidente que hay una diferencia conceptual entre la media y la mediana. Para el lector con ciertas nociones de ingeniería quizá sea de interés que la media de la muestra

www.FreeLibros.me

es el **centroide de los datos** en una muestra. En cierto sentido es el punto en el cual se puede colocar un fulcro (apoyo) para equilibrar un sistema de "pesos", que son las ubicaciones de los datos individuales. Esto se muestra en la figura 1.4 respecto de la muestra "con nitrógeno".

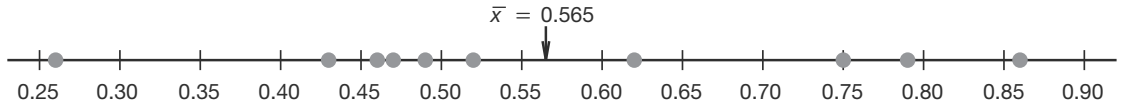


Figura 1.4: Media de la muestra como centroide del peso del tallo con nitrógeno.

En capítulos posteriores la base para el cálculo de $\bar{x}$ es un **estimado** de la **media de la población**. Como antes señalamos, el propósito de la inferencia estadística es obtener conclusiones acerca de las características o **parámetros** y la **estimación** es una característica muy importante de la inferencia estadística.

La mediana y la media pueden ser muy diferentes entre sí. Observe, sin embargo, que en el caso de los datos del peso de los tallos el valor de la media de la muestra para "sin nitrógeno" es bastante similar al valor de la mediana.

Otras medidas de localización

Hay muchos otros métodos para calcular la ubicación del centro de los datos en la muestra. No los trataremos en este momento. Por lo general las alternativas para la media de la muestra se diseñan con el fin de generar valores que representen relación entre la media y la mediana. Rara vez utilizamos alguna de tales medidas. Sin embargo, es aleccionador estudiar una clase de estimadores conocida como **media recortada**, la cual se calcula "quitando" cierto porcentaje de los valores mayores y menores del conjunto. Por ejemplo, la media recortada al 10% se encuentra eliminando tanto el 10% de los valores mayores como el 10% de los menores, y calculando el promedio de los valores restantes. En el caso de los datos del peso de los tallos, eliminaríamos el valor más alto y el más bajo, ya que el tamaño de la muestra es 10 en cada caso. De manera que para el grupo sin nitrógeno la media recortada al 10% está dada por

$$\bar{x}_{\text{rec (10)}} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

y para la media recortada al 10% del grupo con nitrógeno tenemos

$$\bar{x}_{\text{rec (10)}} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Observe que en este caso, como se esperaba, las medias recortadas están cerca tanto de la media como de la mediana para las muestras individuales. Desde luego, el enfoque de la media recortada es menos sensible a los valores extremos que la media de la muestra, pero no tan insensible como la mediana. Además, el método de la media recortada utiliza más información que la mediana de la muestra. Advierta que la mediana de la muestra es, de hecho, un caso especial de la media recortada, en el cual se eliminan todos los datos de la muestra y queda sólo el central o dos observaciones.

www.FreeLibros.me

Ejercicios

1.1 Se registran las siguientes mediciones para el tiempo de secado (en horas) de cierta marca de pintura esmaltada.

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones constituyen una muestra aleatoria simple.

a) ¿Cuál es el tamaño de la muestra anterior?
b) Calcule la media de la muestra para estos datos.
c) Calcule la mediana de la muestra.
d) Grafique los datos utilizando una gráfica de puntos.
e) Calcule la media recortada al 20% para el conjunto de datos anterior.
f) ¿La media muestral para estos datos es más o menos descriptiva como centro de localización, que la media recortada?

1.2 Según la revista *Chemical Engineering*, una propiedad importante de una fibra es su absorción del agua. Se toma una muestra aleatoria de 20 pedazos de fibra de algodón y se mide la absorción de cada uno. Los valores de absorción son los siguientes:

18.71	21.41	20.72	21.81	19.29	22.43	20.17
23.71	19.44	20.50	18.92	20.33	23.00	22.85
19.25	21.77	22.11	19.77	18.04	21.12	

a) Calcule la media y la mediana muestrales para los valores de la muestra anterior.
b) Calcule la media recortada al 10%.
c) Elabore una gráfica de puntos con los datos de la absorción.
d) Si se utilizan sólo los valores de la media, la mediana y la media recortada, ¿hay evidencia de valores extremos en los datos?

1.3 Se utiliza cierto polímero para los sistemas de evacuación de los aviones. Es importante que el polímero sea resistente al proceso de envejecimiento. Se utilizaron veinte especímenes del polímero en un experimento. Diez se asignaron aleatoriamente para exponerse a un proceso de envejecimiento acelerado del lote, el cual implica la exposición a altas temperaturas durante 10 días. Se hicieron las mediciones de resistencia a la tensión de los especímenes y se registraron los siguientes datos sobre resistencia a la tensión en psi.

Sin envejecimiento:	227	222	218	217	225
	218	216	229	228	221
Con envejecimiento:	219	214	215	211	209
	218	203	204	201	205

a) Elabore la gráfica de puntos de los datos.
b) ¿En la gráfica que obtuvo parece que el proceso de envejecimiento tuvo un efecto en la resistencia a la tensión de este polímero? Explique su respuesta.
c) Calcule la resistencia a la tensión de la media de la muestra en las dos muestras.
d) Calcule la mediana de ambas. Analice la similitud o falta de similitud entre la media y la mediana de cada grupo.

1.4 En un estudio realizado por el Departamento de Ingeniería Mecánica del Tecnológico de Virginia se compararon las varillas de acero que abastecen dos compañías diferentes. Se fabricaron diez resortes de muestra con las varillas de metal proporcionadas por cada una de las compañías y se registraron sus medidas de flexibilidad. Los datos son los siguientes:

Compañía A:	9.3	8.8	6.8	8.7	8.5
	6.7	8.0	6.5	9.2	7.0
Compañía B:	11.0	9.8	9.9	10.2	10.1
	9.7	11.0	11.1	10.2	9.6

a) Calcule la media y la mediana de la muestra para los datos de ambas compañías.
b) Grafique los datos para las dos compañías en la misma línea y explique su conclusión respecto de cualquier aparente diferencia entre las dos compañías.

1.5 Veinte hombres adultos de entre 30 y 40 años de edad participaron en un estudio para evaluar el efecto de cierto régimen de salud, que incluye dieta y ejercicio, en el colesterol sanguíneo. Se eligieron aleatoriamente diez para el grupo de control y los otros diez se asignaron para participar en el régimen como el grupo de tratamiento durante un periodo de seis meses. Los siguientes datos muestran la reducción en el colesterol que experimentaron en ese periodo los 20 sujetos:

Grupo de control:	7	3	−4	14	2
	5	22	−7	9	5
Grupo de tratamiento:	−6	5	9	4	4
	12	37	5	3	3

a) Elabore una gráfica de puntos con los datos de ambos grupos en la misma gráfica.
b) Calcule la media, la mediana y la media recortada al 10% para ambos grupos.
c) Explique por qué la diferencia en las medias sugiere una conclusión acerca del efecto del régimen, en tanto que la diferencia en las medianas o las medias recortadas sugiere una conclusión diferente.

1.6 La resistencia a la tensión del caucho de silicio se considera una función de la temperatura de vulcanizado. Se llevó a cabo un estudio en el que se prepararon muestras de 12 especímenes del caucho utilizando temperaturas de vulcanizado de 20°C y 45°C. Los siguientes

www.FreeLibros.me

datos presentan los valores de resistencia a la tensión en megapascales.

20°C:	2.07	2.14	2.22	2.03	2.21	2.03
	2.05	2.18	2.09	2.14	2.11	2.02
45°C:	2.52	2.15	2.49	2.03	2.37	2.05
	1.99	2.42	2.08	2.42	2.29	2.01

a) Elabore una gráfica de puntos con los datos, tanto de los valores de resistencia a la tensión a temperatura alta como los de a temperatura baja.

b) Calcule la resistencia a la tensión media muestral para ambas muestras.

c) Al observar la gráfica, ¿le parece que la temperatura de vulcanizado influye en la resistencia a la tensión? Explique su respuesta.

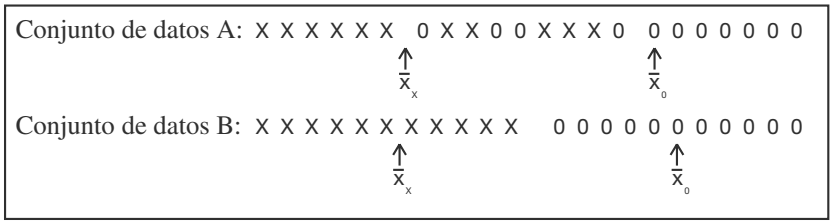
d) ¿En qué otra cosa, al parecer, influye el incremento en la temperatura de vulcanizado? Explique su respuesta.

1.4 Medidas de variabilidad

La variabilidad de una muestra desempeña un papel importante en el análisis de datos. La variabilidad de procesos y productos es un hecho real en los sistemas científicos y de ingeniería: el control o la reducción de la variabilidad de un proceso a menudo es una fuente de mayores dificultades. Cada vez más ingenieros y administradores de procesos están aprendiendo que la calidad del producto y, como resultado, las utilidades que se derivan de los productos manufacturados es, con mucho, una función de la **variabilidad del proceso**. En consecuencia, gran parte de los capítulos 9 a 15 se dedica al análisis de datos y a los procedimientos de modelado en los que la variabilidad de la muestra desempeña un papel significativo. Incluso en problemas pequeños de análisis de datos el éxito de un método estadístico específico podría depender de la magnitud de la variabilidad entre las observaciones en la muestra. Las medidas de ubicación en una muestra no brindan un resumen adecuado de la naturaleza de un conjunto de datos. Considere el ejemplo 1.2, en el que no podemos concluir que el uso del nitrógeno aumenta el crecimiento sin tomar en cuenta la variabilidad de la muestra.

Aunque los detalles del análisis de este tipo de conjuntos de datos se estudiarán en el capítulo 9, a partir de la figura 1.1 debería quedar claro que la variabilidad entre las observaciones sin nitrógeno y la variabilidad entre las observaciones con nitrógeno tiene, desde luego, alguna consecuencia. De hecho, parece que la variabilidad dentro de la muestra con nitrógeno es mayor que la de la muestra sin nitrógeno. Quizás haya algo acerca de la inclusión del nitrógeno que no sólo incrementa el peso de los tallos ($\bar{x}$ de 0.565 gramos en comparación con una $\bar{x}$ de 0.399 gramos para la muestra sin nitrógeno), sino que también incrementa la variabilidad en el peso de los tallos (es decir, provoca que el peso de los tallos sea más inconsistente).

Por ejemplo, compare los dos conjuntos de datos de abajo. Cada uno contiene dos muestras y la diferencia en las medias es aproximadamente la misma para ambas, aunque el conjunto de datos B parece proporcionar un contraste mucho más claro entre las dos poblaciones de las que se tomaron las muestras. Si el propósito de tal experimento es detectar la diferencia entre las dos poblaciones, esto se logra en el caso del conjunto de datos B. Sin embargo, en el conjunto de datos A la amplia variabilidad *dentro* de las dos muestras ocasiona dificultad. De hecho, no es claro que haya una diferencia *entre* las dos poblaciones.



www.FreeLibros.me

Rango y desviación estándar de la muestra

Así como hay muchas medidas de tendencia central o de localización, hay muchas medidas de dispersión o variabilidad. Quizá la más simple sea el **rango de la muestra** $X_{\text{máx}} - X_{\text{mín}}$. El rango puede ser muy útil y se examina con amplitud en el capítulo 17 sobre *control estadístico de calidad.* La medida muestral de dispersión que se utiliza más a menudo es la **desviación estándar de la muestra**. Nuevamente denotemos con x_1, $x_2, \ldots, x_n$ los valores de la muestra.

Definición 1.3: La **varianza de la muestra**, denotada con s^2, está dada por

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}.$$

La **desviación estándar de la muestra**, denotada con s, es la raíz cuadrada positiva de s^2, es decir,

$$s = \sqrt{s^2}.$$

Para el lector debería quedar claro que la desviación estándar de la muestra es, de hecho, una medida de variabilidad. Una variabilidad grande en un conjunto de datos produce valores relativamente grandes de $(x - \bar{x})^2$ y, por consiguiente, una varianza muestral grande. La cantidad $n - 1$ a menudo se denomina **grados de libertad asociados con la varianza** estimada. En este ejemplo sencillo los grados de libertad representan el número de piezas de información independientes disponibles para calcular la variabilidad. Por ejemplo, suponga que deseamos calcular la varianza de la muestra y la desviación estándar del conjunto de datos (5, 17, 6, 4). El promedio de la muestra es $\bar{x} = 8$. El cálculo de la varianza implica:

$$(5-8)^2 + (17-8)^2 + (6-8)^2 + (4-8)^2 = (-3)^2 + 9^2 + (-2)^2 + (-4)^2.$$

Las cantidades dentro de los paréntesis suman cero. En general, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (véase el ejercicio 1.16 de la página 31). Entonces, el cálculo de la varianza de una muestra no implica n **desviaciones cuadradas independientes** de la media $\bar{x}$. De hecho, como el último valor de $x - \bar{x}$ es determinado por los primeros $n - 1$ valores, decimos que éstas son $n - 1$ "piezas de información" que produce s^2. Por consiguiente, hay $n - 1$ grados de libertad en vez de n grados de libertad para calcular la varianza de una muestra.

Ejemplo 1.4: En un ejemplo que se estudia ampliamente en el capítulo 10, un ingeniero se interesa en probar el "sesgo" en un medidor de pH. Los datos se recaban con el medidor mediante la medición del pH de una sustancia neutra (pH = 7.0). Se toma una muestra de tamaño 10 y se obtienen los siguientes resultados:

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08.

La media de la muestra $\bar{x}$ está dada por

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \ldots + 7.08}{10} = 7.0250.$$

www.FreeLibros.me

La varianza de la muestra s^2 está dada por

$$s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 + \cdots + (7.08 - 7.025)^2] = 0.001939.$$

Como resultado, la desviación estándar de la muestra está dada por

$$s = \sqrt{0.001939} = 0.044.$$

Así que la desviación estándar de la muestra es 0.0440 con $n - 1 = 9$ grados de libertad. ∎

Unidades para la desviación estándar y la varianza

A partir de la definición 1.3 debería ser evidente que la varianza es una medida de la desviación cuadrática promedio de la media $\bar{x}$. Empleamos el término *desviación cuadrática promedio* aun cuando la definición utilice una división entre $n - 1$ grados de libertad en vez de n. Desde luego, si n es grande, la diferencia en el denominador es inconsecuente. Por lo tanto, la varianza de la muestra tiene unidades que son el cuadrado de las unidades en los datos observados; aunque la desviación estándar de la muestra se encuentra en unidades lineales. Considere los datos del ejemplo 1.2. Los pesos del tallo se miden en gramos. Como resultado, las desviaciones estándar de la muestra están en gramos y las varianzas se miden en gramos^2. De hecho, las desviaciones estándar individuales son 0.0728 gramos para el caso sin nitrógeno y 0.1867 gramos para el grupo con nitrógeno. Observe que la desviación estándar en verdad indica una variabilidad mucho más grande en la muestra con nitrógeno. Esta condición se destaca en la figura 1.1.

¿Cuál es la medida de variabilidad más importante?

Como indicamos antes, el rango de la muestra tiene aplicaciones en el área del control estadístico de la calidad. Quizás el lector considere que es redundante utilizar la varianza de la muestra y la desviación estándar de la muestra. Ambas medidas reflejan el mismo concepto en la variabilidad de la medición, pero la desviación estándar de la muestra mide la variabilidad en unidades lineales; en tanto que la varianza muestral se mide en unidades cuadradas. Ambas desempeñan papeles importantes en el uso de los métodos estadísticos. Mucho de lo que se logra en el contexto de la inferencia estadística implica la obtención de conclusiones acerca de las características de poblaciones. Entre tales características son constantes los denominados **parámetros de la población**. Dos parámetros importantes son la **media de la población** y la **varianza de la población**. La varianza de la muestra desempeña un papel explícito en los métodos estadísticos que se utilizan para obtener inferencias sobre la varianza de la población. La desviación estándar de la muestra desempeña un papel importante, junto con la media de la muestra, en las inferencias que se realizan acerca de la media de la población. En general, la varianza se considera más en la teoría inferencial, mientras que la desviación estándar se utiliza más en aplicaciones.

www.FreeLibros.me

Ejercicios

1.7 Considere los datos del tiempo de secado del ejercicio 1.1 de la página 13. Calcule la varianza de la muestra y la desviación estándar de la muestra.

1.8 Calcule la varianza de la muestra y la desviación estándar para los datos de absorción del agua del ejercicio 1.2 de la página 13.

1.9 El ejercicio 1.3 de la página 13 presentó datos de resistencia a la tensión de dos muestras, una en la que los especímenes se expusieron a un proceso de envejecimiento y otra en la que no se efectuó tal proceso en los especímenes.

a) Calcule la varianza de la muestra, así como su desviación estándar, en cuanto a la resistencia a la tensión en ambas muestras.

b) ¿Parece haber alguna evidencia de que el envejecimiento afecta la variabilidad en la resistencia a la tensión? (Véase también la gráfica para el ejercicio 1.3 de la página 13).

1.10 Para los datos del ejercicio 1.4 de la página 13 calcule tanto la media como la varianza de la "flexibilidad" para las compañías A y B. ¿Parece que hay una diferencia de flexibilidad entre la compañía A y la compañía B?

1.11 Considere los datos del ejercicio 1.5 de la página 13. Calcule la varianza de la muestra y la desviación estándar de la muestra para ambos grupos: el de tratamiento y el de control.

1.12 Para el ejercicio 1.6 de la página 13 calcule la desviación estándar muestral de la resistencia a la tensión para las muestras, de forma separada para ambas temperaturas. ¿Parece que un incremento en la temperatura influye en la variabilidad de la resistencia a la tensión? Explique su respuesta.

1.5 Datos discretos y continuos

La inferencia estadística a través del análisis de estudios observacionales o de diseños experimentales se utiliza en muchas áreas científicas. Los datos reunidos pueden ser **discretos** o **continuos**, según el área de aplicación. Por ejemplo, un ingeniero químico podría estar interesado en un experimento que lo lleve a condiciones en que se maximice la producción. Aquí, por supuesto, la producción se expresaría en porcentaje, o gramos/libra, medida en un continuo. Por otro lado, un toxicólogo que realice un experimento de combinación de fármacos quizás encuentre datos que son binarios por naturaleza (es decir, el paciente responde o no lo hace).

En la teoría de la probabilidad se hacen distinciones importantes entre datos discretos y continuos que nos permiten hacer inferencias estadísticas. Con frecuencia las aplicaciones de la inferencia estadística se encuentran cuando se trabaja con *datos por conteo*. Por ejemplo, un ingeniero podría estar interesado en estudiar el número de partículas radiactivas que pasan a través de un contador en, digamos, 1 milisegundo. Al personal responsable de la eficiencia de una instalación portuaria podría interesarle conocer las características del número de buques petroleros que llegan diariamente a cierta ciudad portuaria. En el capítulo 5 se examinarán varios escenarios diferentes que conducen a distintas formas de manejo de los datos para situaciones con datos por conteo.

Incluso en esta fase inicial del texto se debería poner especial atención a algunos detalles que se asocian con datos binarios. Son muchas las aplicaciones que requieren el análisis estadístico de datos binarios. Con frecuencia la medición que se utiliza en el análisis es la *proporción muestral*. En efecto, la situación binaria implica dos categorías. Si en los datos hay n unidades y x se define como el número que cae en la categoría 1, entonces $n - x$ cae en la categoría 2. Así, x/n es la proporción muestral en la categoría 1 y $1 - x/n$ es la proporción muestral en la categoría 2. En la aplicación biomédica, por ejemplo, 50 pacientes representarían las unidades de la muestra y si, después de que se les suministra el medicamento, 20 de los 50 experimentaran mejoría en sus malestares estomacales (que son comunes en los 50), entonces $\frac{20}{50} = 0.4$ sería la proporción muestral

www.FreeLibros.me

para la cual el medicamento tuvo éxito, y $1 - 0.4 = 0.6$ sería la proporción muestral para la cual el fármaco no tuvo éxito. En realidad, la medición numérica fundamental para datos binarios por lo general se denota con 0 o con 1. Éste es el caso de nuestro ejemplo médico, en el que un resultado exitoso se denota con un 1 y uno no exitoso con un 0. Entonces, la proporción muestral es en realidad una media muestral de unos y ceros. Para la categoría de éxitos,

$$\frac{x_1 + x_2 + \cdots + x_{50}}{50} = \frac{1 + 1 + 0 + \cdots + 0 + 1}{50} = \frac{20}{50} = 0.4.$$

¿Qué clases de problemas se resuelven en situaciones con datos binarios?

Los tipos de problemas que enfrentan científicos e ingenieros que usan datos binarios no son muy difíciles, a diferencia de aquellos en los que las mediciones de interés son las continuas. Sin embargo, se utilizan técnicas diferentes debido a que las propiedades estadísticas de las proporciones muestrales son bastante diferentes de las medias muestrales que resultan de los promedios tomados de poblaciones continuas. Considere los datos del ejemplo en el ejercicio 1.6 de la página 13. El problema estadístico subyacente en este caso se enfoca en si una intervención, digamos un incremento en la temperatura de vulcanizado, alterará la resistencia a la tensión de la media de la población que se asocia con el proceso del caucho de silicio. Por otro lado, en el área de control de calidad, suponga que un fabricante de neumáticos para automóvil informa que en un embarque con 5000 neumáticos, seleccionados aleatoriamente del proceso, hay 100 defectuosos. Aquí la proporción muestral es $\frac{100}{5000} = 0.02$. Luego de realizar un cambio en el proceso diseñado para reducir los neumáticos defectuosos, se toma una segunda muestra de 5000 y se encuentran 90 defectuosos. La proporción muestral se redujo a $\frac{90}{5000} = 0.018$. Entonces, surge una pregunta: "¿La disminución en la proporción muestral de 0.02 a 0.018 es suficiente para sugerir una mejoría real en la proporción de la población?" En ambos casos se requiere el uso de las propiedades estadísticas de los promedios de la muestra: uno de las muestras de poblaciones continuas y el otro de las muestras de poblaciones discretas (binarias). En ambos casos la media de la muestra es un **estimado** de un parámetro de la población, una media de la población en el primer caso (es decir, la media de la resistencia a la tensión) y una proporción de la población (o sea, la proporción de neumáticos defectuosos en la población) en el segundo caso. Así que aquí tenemos estimados de la muestra que se utilizan para obtener conclusiones científicas respecto de los parámetros de la población. Como indicamos en la sección 1.3, éste es el tema general en muchos problemas prácticos en los que se usa la inferencia estadística.

1.6 Modelado estadístico, inspección científica y diagnósticos gráficos

A menudo, el resultado final de un análisis estadístico es la estimación de los parámetros de un **modelo postulado.** Éste es un proceso natural para los científicos y los ingenieros, ya que con frecuencia usan modelos. Un modelo estadístico no es determinista, es más bien un modelo que conlleva algunos aspectos probabilísticos. A menudo una forma de modelo es la base de las **suposiciones** que hace el analista. En el ejemplo 1.2 el científico podría desear determinar, a través de la información de la muestra, algún nivel de distinción entre las poblaciones tratadas con nitrógeno y las poblaciones no tratadas. El análisis podría requerir cierto modelo para los datos; por ejemplo, que las dos muestras

www.FreeLibros.me

provengan de **distribuciones normales** o **gaussianas**. Véase el capítulo 6 para el estudio de la distribución normal.

Es evidente que quienes utilizan métodos estadísticos no pueden generar la información o los datos experimentales suficientes para describir a la totalidad de la población. Pero es frecuente que se utilicen los conjuntos de datos para aprender sobre ciertas propiedades de la población. Los científicos y los ingenieros están acostumbrados a manejar conjuntos de datos. Debería ser obvia la importancia de describir o *resumir* la naturaleza de los conjuntos de datos. Con frecuencia el resumen gráfico de un conjunto de datos puede proporcionar información sobre el sistema del que se obtuvieron los datos. Por ejemplo, en las secciones 1.1 y 1.3 mostramos gráficas de puntos.

En esta sección se estudia con detalle el papel del muestreo y de la graficación de los datos para mejorar la **inferencia estadística**. Nos limitamos a presentar algunas gráficas sencillas, pero a menudo efectivas, que complementan el estudio de poblaciones estadísticas.

Diagrama de dispersión

A veces el modelo postulado puede tener una forma algo más compleja. Por ejemplo, considere a un fabricante de textiles que diseña un experimento en donde se producen especímenes de tela que contienen diferentes porcentajes de algodón. Considere los datos de la tabla 1.3.

Tabla 1.3: Resistencia a la tensión

Porcentaje del algodón	Resistencia a la tensión
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10

Se fabrican cinco especímenes de tela para cada uno de los cuatro porcentajes de algodón. En este caso tanto el modelo para el experimento como el tipo de análisis que se utiliza deberían tomar en cuenta el objetivo del experimento y los insumos importantes del científico textil. Algunas gráficas sencillas podrían mostrar la clara distinción entre las muestras. Véase la figura 1.5; las medias y la variabilidad muestrales se describen bien en el diagrama de dispersión. El objetivo de este experimento podría ser simplemente determinar cuáles porcentajes de algodón son verdaderamente distintos de los otros. En otras palabras, como en el caso de los datos con nitrógeno y sin nitrógeno, ¿para cuáles porcentajes de algodón existen diferencias claras entre las poblaciones o, de forma más específica, entre las medias de las poblaciones? En este caso quizás un modelo razonable es que cada muestra proviene de una distribución normal. Aquí el objetivo es muy semejante al de los datos con nitrógeno y sin nitrógeno, excepto que se incluyen más muestras. El formalismo del análisis implica nociones de prueba de hipótesis, los cuales se examinarán en el capítulo 10. A propósito, quizás este formalismo no sea necesario a la luz del diagrama de diagnóstico. Pero, ¿describe éste el objetivo real del experimento y, por consiguiente, el enfoque adecuado para el análisis de datos? Es probable que el científico anticipe la existencia de una *resistencia a la tensión máxima de la media de la población* en el rango de concentración de algodón en el experimento. Aquí el análisis de los datos debería girar en torno a un tipo diferente de modelo, es decir, uno

www.FreeLibros.me

que postule un tipo de estructura que relacione la resistencia a la tensión de la media de la población con la concentración de algodón. En otras palabras, un modelo se puede escribir como

$$\mu_{t,c} = \beta_0 + \beta_1 C + \beta_2 C^2,$$

en donde $\mu_{t,c}$ es la resistencia a la tensión de la media de la población, que varía con la cantidad de algodón en el producto C. La implicación de este modelo es que, para un nivel fijo de algodón, hay una población de mediciones de resistencia a la tensión y la media de la población es $\mu_{t,c}$. Este tipo de modelo, que se denomina **modelo de regresión**, se estudiará en los capítulos 11 y 12. La forma funcional la elige el científico. A veces el análisis de datos puede sugerir que se cambie el modelo. Entonces el analista de datos "considera" un modelo que se pueda alterar después de hacer cierto análisis. El uso de un modelo empírico va acompañado por la **teoría de estimación**, donde β_0, β_1 y β_2 se estiman a partir de los datos. Además, la inferencia estadística se puede, entonces, utilizar para determinar lo adecuado del modelo.

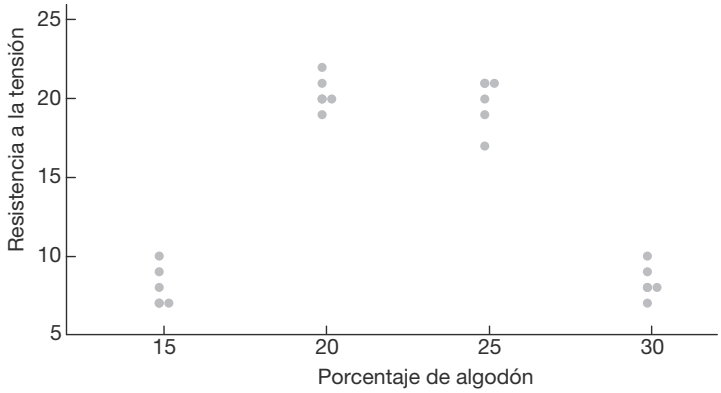


Figura 1.5: Diagrama de dispersión de la resistencia a la tensión y los porcentajes de algodón.

Aquí se hacen evidentes dos puntos de las dos ilustraciones de datos: 1) el tipo de modelo que se emplea para describir los datos a menudo depende del objetivo del experimento, y 2) la estructura del modelo debería aprovechar el insumo científico no estadístico. La selección de un modelo representa una **suposición fundamental** sobre la que se basa la inferencia estadística resultante. A lo largo del libro se hará evidente la importancia que las gráficas pueden llegar a tener. A menudo las gráficas ilustran información que permite que los resultados de la inferencia estadística formal se comuniquen mejor al científico o al ingeniero. A veces las gráficas o el **análisis exploratorio de los datos** pueden enseñar al analista información que no se obtiene del análisis formal. Casi cualquier análisis formal requiere suposiciones que se desarrollan a partir del modelo de datos. Las gráficas pueden resaltar la **violación de suposiciones** que de otra forma no se notarían. A lo largo del libro las gráficas se utilizarán de manera extensa para complementar el análisis formal de los datos. En las siguientes secciones se presentan algunas herramientas gráficas que son útiles para el análisis exploratorio o descriptivo de los datos.

www.FreeLibros.me

Diagrama de tallo y hojas

Los datos estadísticos obtenidos de poblaciones grandes pueden ser muy útiles para estudiar el comportamiento de la distribución si se presentan en una combinación tabular y gráfica conocida como **diagrama de tallo y hojas**.

Para ejemplificar la elaboración de un diagrama de tallo y hojas considere los datos de la tabla 1.4, que especifican la "vida" de 40 baterías para automóvil similares, registradas al décimo de año más cercano. Las baterías se garantizan por tres años. Comience por dividir cada observación en dos partes: una para el tallo y otra para las hojas, de manera que el tallo represente el dígito entero que antecede al decimal y la hoja corresponda a la parte decimal del número. En otras palabras, para el número 3.7 el dígito 3 se designa al tallo y el 7 a la hoja. Para nuestros datos los cuatro tallos 1, 2, 3 y 4 se listan verticalmente del lado izquierdo de la tabla 1.5, en tanto que las hojas se registran en el lado derecho correspondiente al valor del tallo adecuado. Entonces, la hoja 6 del número 1.6 se registra enfrente del tallo 1; la hoja 5 del número 2.5 enfrente del tallo 2; y así sucesivamente. El número de hojas registrado junto a cada tallo se anota debajo de la columna de frecuencia.

Tabla 1.4: Vida de las baterías para automóvil

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Tabla 1.5: Diagrama de tallo y hojas de la vida de las baterías

Tallo	**Hoja**	**Frecuencia**
1	69	2
2	25669	5
3	00111122233344455677889 9	25
4	11234577	8

El diagrama de tallo y hojas de la tabla 1.5 contiene sólo cuatro tallos y, en consecuencia, no ofrece una representación adecuada de la distribución. Para solucionar este problema es necesario aumentar el número de tallos en nuestro diagrama. Una manera sencilla de hacerlo consiste en escribir dos veces cada valor del tallo y después registrar las hojas 0, 1, 2, 3 y 4 enfrente del valor del tallo adecuado, donde aparezca por primera vez; y las hojas 5, 6, 7, 8 y 9 enfrente de este mismo valor del tallo, donde aparece la segunda vez. El diagrama doble de tallo y hojas modificado se ilustra en la tabla 1.6, donde los tallos que corresponden a las hojas 0 a 4 fueron codificados con el símbolo $\star$ y los tallos correspondientes a las hojas 5 a 9 con el símbolo $\bullet$.

En cualquier problema dado debemos decidir cuáles son los valores del tallo adecuados. Esta decisión se toma hasta cierto punto de manera arbitraria, aunque debemos guiarnos por el tamaño de nuestra muestra. Por lo general elegimos entre 5 y 20 tallos. Cuanto más pequeña sea la cantidad de datos disponibles, más pequeña será nuestra elección del número de tallos. Por ejemplo, si los datos constan de números del 1 al 21,

www.FreeLibros.me

los cuales representan el número de personas en la fila de una cafetería en 40 días laborables seleccionados al azar, y elegimos un diagrama doble de tallo y hojas, los tallos serían 0⋆, 0·, 1⋆, 1· y 2⋆, de manera que la observación de 1 más pequeña tiene tallo 0⋆ y hoja 1, el número 18 tiene tallo 1· y hoja 8, y la observación de 21 más grande tiene tallo 2⋆ y hoja 1. Por otro lado, si los datos constan de números de $18,800 a $19,600, que representan las mejores ventas posibles de 100 automóviles nuevos, obtenidos de cierto concesionario, y elegimos un diagrama sencillo de tallo y hojas, los tallos serían 188, 189, 190, …, 196 y las hojas contendrían ahora dos dígitos cada una. Un automóvil que se vende en $19,385 tendría un valor de tallo de 193 y 85 en los dos dígitos de la hoja. En el diagrama de tallo y hojas, las hojas de dígitos múltiples que pertenecen al mismo tallo por lo regular están separadas por comas. En los datos generalmente se ignoran los puntos decimales cuando todos los números a la derecha del punto decimal representan hojas, como en el caso de las tablas 1.5 y 1.6. Sin embargo, si los datos constaran de números que van de 21.8 a 74.9, podríamos elegir los dígitos 2, 3, 4, 5, 6 y 7 como los tallos, de manera que un número como 48.3 tendría un valor de tallo de 4 y un valor de hoja de 8.3.

Tabla 1.6: Diagrama doble de tallo y hojas para la vida de las baterías

Tallo	Hoja	Frecuencia
1·	69	2
2⋆	2	1
2·	5669	4
3⋆	001111222333444	15
3·	5567778899	10
4⋆	11234	5
4·	577	3

El diagrama de tallo y hojas representa una manera eficaz de resumir los datos. Otra forma consiste en el uso de la **distribución de frecuencias**, donde los datos, agrupados en diferentes clases o intervalos, se pueden construir contando las hojas que pertenecen a cada tallo y considerando que cada tallo define un intervalo de clase. En la tabla 1.5 el tallo 1 con 2 hojas define el intervalo 1.0-1.9, que contiene 2 observaciones; el tallo 2 con 5 hojas define el intervalo 2.0-2.9, que contiene 5 observaciones; el tallo 3 con 25 hojas define el intervalo 3.0-3.9, con 25 observaciones; y el tallo 4 con 8 hojas define el intervalo 4.0-4.9, que contiene 8 observaciones. Para el diagrama doble de tallo y hojas de la tabla 1.6 los tallos definen los siete intervalos de clase 1.5-1.9, 2.0-2.4, 2.5-2.9, 3.0-3.4, 3.5-3.9, 4.0-4.4 y 4.5-4.9, con frecuencias 2, 1, 4, 15, 10, 5 y 3, respectivamente.

Histograma

Al dividir cada frecuencia de clase entre el número total de observaciones, obtenemos la proporción del conjunto de observaciones en cada una de las clases. Una tabla que lista las frecuencias relativas se denomina **distribución de frecuencias relativas**. En la tabla 1.7 se presenta la distribución de frecuencias relativas para los datos de la tabla 1.4, que muestra los puntos medios de cada intervalo de clase.

La información que brinda una distribución de frecuencias relativas en forma tabular es más fácil de entender si se presenta en forma gráfica. Con los puntos medios de

www.FreeLibros.me

Tabla 1.7: Distribución de frecuencias relativas de la vida de las baterías

Intervalo de clase	Punto medio de la clase	Frecuencia, f	Frecuencia relativa
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

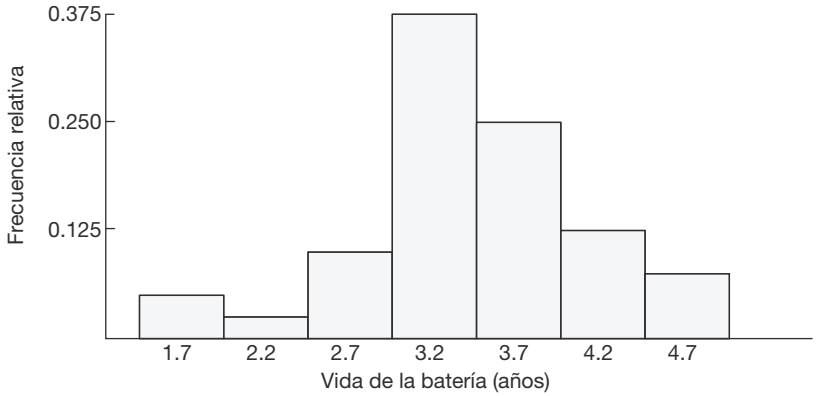


Figura 1.6: Histograma de frecuencias relativas.

cada intervalo y las frecuencias relativas correspondientes construimos un **histograma de frecuencias relativas** (figura 1.6).

Muchas distribuciones de frecuencias continuas se pueden representar gráficamente mediante la curva en forma de campana característica de la figura 1.7. Herramientas gráficas como las de las figuras 1.6 y 1.7 ayudan a comprender la naturaleza de la población. En los capítulos 5 y 6 examinaremos una propiedad de la población que se conoce como **distribución**. Aunque más adelante en este texto se proporcionará una definición más precisa de una distribución o de una **distribución de probabilidad**, aquí podemos visualizarla como la que se podría haber visto en el límite de la figura 1.7 cuando el tamaño de la muestra aumentara.

Se dice que una distribución es **simétrica** si se puede doblar a lo largo de un eje vertical de manera que ambos lados coincidan. Si una distribución carece de simetría respecto de un eje vertical, se dice que está **sesgada**. La distribución que se ilustra en la figura 1.8a se dice que está sesgada a la derecha porque tiene una cola derecha larga y una cola izquierda mucho más corta. En la figura 1.8b observamos que la distribución es simétrica; mientras que en la figura 1.8c está sesgada a la izquierda.

Al girar un diagrama de tallo y hojas en dirección contraria a la de las manecillas del reloj en un ángulo de 90°, vemos que las columnas de hojas que resultan forman una imagen parecida a un histograma. Por lo tanto, si nuestro objetivo principal al observar los datos es determinar la forma general o la forma de la distribución, rara vez será necesario construir un histograma de frecuencias relativas.

www.FreeLibros.me

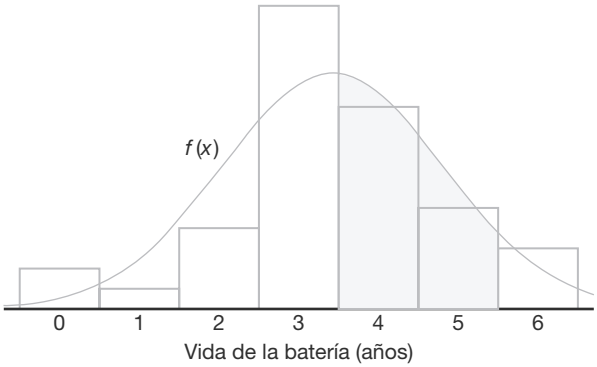


Figura 1.7: Estimación de la distribución de frecuencias.

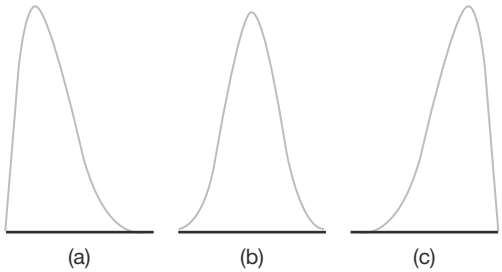


Figura 1.8: Sesgo de los datos.

Gráfica de caja y bigote o gráfica de caja

Otra presentación que es útil para reflejar propiedades de una muestra es la **gráfica de caja y bigote**, la cual encierra el *rango intercuartil* de los datos en una caja que contiene la mediana representada. El rango intercuartil tiene como extremos el percentil 75 (cuartil superior) y el percentil 25 (cuartil inferior). Además de la caja se prolongan "bigotes", que indican las observaciones alejadas en la muestra. Para muestras razonablemente grandes la presentación indica el centro de localización, la variabilidad y el grado de asimetría.

Además, una variación denominada **gráfica de caja** puede ofrecer al observador información respecto de cuáles observaciones son **valores extremos**. Los valores extremos son observaciones que se consideran inusualmente alejadas de la masa de datos. Existen muchas pruebas estadísticas diseñadas para detectar este tipo de valores. Técnicamente se puede considerar que un valor extremo es una observación que representa un "evento raro" (existe una probabilidad pequeña de obtener un valor que esté lejos de la masa de datos). El concepto de valores extremos volverá a surgir en el capítulo 12 en el contexto del análisis de regresión.

www.FreeLibros.me

La información visual en las gráficas de caja y bigote o en las de caja no intenta ser una prueba formal de valores extremos, más bien se considera una herramienta de diagnóstico. Aunque la determinación de cuáles observaciones son valores extremos varía de acuerdo con el tipo de software que se emplee, un procedimiento común para determinarlo consiste en utilizar un **múltiplo del rango intercuartil**. Por ejemplo, si la distancia desde la caja excede 1.5 veces el rango intercuartil (en cualquier dirección), la observación se podría considerar un valor extremo.

Ejemplo 1.5: Se midió el contenido de nicotina en una muestra aleatoria de 40 cigarrillos. Los datos se presentan en la tabla 1.8.

Tabla 1.8: Valores de nicotina para el ejemplo 1.5

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

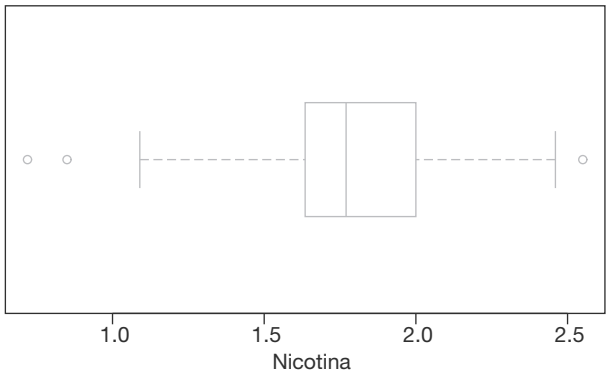


Figura 1.9: Gráfica de caja y bigote para el ejemplo 1.5.

La figura 1.9 muestra la gráfica de caja y bigote de los datos, la cual describe las observaciones 0.72 y 0.85 como valores extremos moderados en la cola inferior; en tanto que la observación 2.55 es un valor extremo moderado en la cola superior. En este ejemplo el rango intercuartil es 0.365, y 1.5 veces el rango intercuartil es 0.5475. Por otro lado, la figura 1.10 presenta un diagrama de tallo y hojas. ◢

Ejemplo 1.6: Considere los datos de la tabla 1.9, que constan de 30 muestras que miden el grosor de las "asas" de latas de pintura (véase el trabajo de Hogg y Ledolter de 1992 en la bibliografía). La figura 1.11 describe una gráfica de caja y bigote para este conjunto asimétrico de datos. Observe que el bloque izquierdo es considerablemente más grande que el bloque de la derecha. La mediana es 35. El cuartil inferior es 31, mientras que el superior es 36. Advierta también que la observación alejada de la derecha está más lejos de la caja que la observación extrema de la izquierda. No hay valores extremos en este conjunto de datos. ◢

www.FreeLibros.me

```
El punto decimal se encuentra 1 dígito(s) a la izquierda de I
 7 | 2
 8 | 5
 9 |
10 | 9
11 |
12 | 4
13 | 7
14 | 07
15 | 18
16 | 3447899
17 | 045599
18 | 2568
19 | 0237
20 | 389
21 | 17
22 | 8
23 | 17
24 | 6
25 | 5
```

Figura 1.10: Diagrama de tallo y hojas para los datos de nicotina.

Tabla 1.9: Datos para el ejemplo 1.6

Muestra	Mediciones	Muestra	Mediciones
1	29 36 39 34 34	16	35 30 35 29 37
2	29 29 28 32 31	17	40 31 38 35 31
3	34 34 39 38 37	18	35 36 30 33 32
4	35 37 33 38 41	19	35 34 35 30 36
5	30 29 31 38 29	20	35 35 31 38 36
6	34 31 37 39 36	21	32 36 36 32 36
7	30 35 33 40 36	22	36 37 32 34 34
8	28 28 31 34 30	23	29 34 33 37 35
9	32 36 38 38 35	24	36 36 35 37 37
10	35 30 37 35 31	25	36 30 35 33 31
11	35 30 35 38 35	26	35 30 29 38 35
12	38 34 35 35 31	27	35 36 30 34 36
13	34 35 33 30 34	28	35 30 36 29 35
14	40 35 34 33 35	29	38 36 35 31 31
15	34 35 38 35 30	30	30 34 40 28 30

Existen otras formas en las que las gráficas de caja y bigote, y otras presentaciones gráficas, pueden ayudar al analista. Las muestras múltiples se pueden comparar de forma gráfica. Los diagramas de los datos pueden sugerir relaciones entre las variables y las gráficas ayudan a detectar anomalías u observaciones extremas en las muestras.

Existen otros tipos diferentes de diagramas y herramientas gráficas, los cuales se estudiarán en el capítulo 8 después de presentar otros detalles teóricos.

www.FreeLibros.me

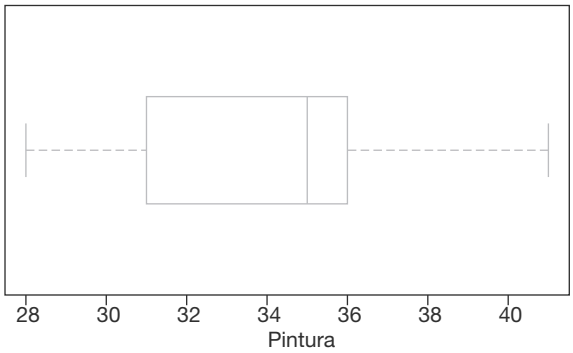


Figura 1.11: Gráfica de caja y bigote del grosor de las "asas" de latas de pintura.

Otras características distintivas de una muestra

Hay características de la distribución o de la muestra, además de las medidas del centro de localización y variabilidad, que definen aún más su naturaleza. Por ejemplo, en tanto que la mediana divide los datos (o su distribución) en dos partes, existen otras medidas que dividen partes o segmentos de la distribución que pueden ser muy útiles. Una separación en cuatro partes se hace mediante *cuartiles*, donde el tercer cuartil separa el cuarto (25%) superior del resto de los datos, el segundo cuartil es la mediana y el primer cuartil separa el cuarto (25%) inferior del resto de los datos. La distribución puede dividirse incluso más detalladamente calculando los percentiles. Tales cantidades dan al analista una noción de las denominadas *colas* de la distribución (es decir, los valores que son relativamente extremos, ya sean pequeños o grandes). Por ejemplo, el percentil 95 separa el 5% superior del 95% inferior. Para los extremos en la parte inferior o *cola inferior* de la distribución prevalecen definiciones similares. El primer percentil separa el 1% inferior del resto de la distribución. El concepto de percentiles desempeñará un papel significativo en buena parte de lo que estudiaremos en los siguientes capítulos.

1.7 Tipos generales de estudios estadísticos: diseño experimental, estudio observacional y estudio retrospectivo

En las siguientes secciones destacaremos el concepto de muestreo de una población y el uso de los métodos estadísticos para aprender o quizá para reafirmar la información relevante acerca de una población. La información que se busca y que se obtiene mediante el uso de tales métodos estadísticos a menudo influye en la toma de decisiones, así como en la resolución de problemas en diversas áreas importantes de ingeniería y científicas. Como ilustración, el ejemplo 1.3 describe un experimento sencillo, en el cual los resultados brindan ayuda para determinar los tipos de condiciones en los que no se recomienda utilizar una aleación de aluminio específica que podría ser muy vulnerable a la corrosión. Los resultados serían útiles no sólo para quienes fabrican la aleación, sino también para los clientes que consideren adquirirla. Este caso, y muchos otros que se incluyen en los capítulos 13 a 15, resaltan el concepto de condiciones experimentales diseñadas o controladas (combinaciones de condiciones de recubrimiento y humedad), que son de interés para aprender sobre algunas características o mediciones (nivel de corrosión) que

www.FreeLibros.me

surgen de tales condiciones. En las mediciones de la corrosión se emplean métodos estadísticos que utilizan tanto medidas de tendencia central como de variabilidad. Como usted verá más adelante en este texto, tales métodos con frecuencia nos guían hacia un modelo estadístico como el que se examinó en la sección 1.6. En este caso el modelo se puede usar para estimar (o predecir) las medidas de la corrosión como una función de la humedad y el tipo de recubrimiento utilizado. De nuevo, para desarrollar este tipo de modelos es muy útil emplear las estadísticas descriptivas que destacan las medidas de tendencia central y de variabilidad.

La información que se ofrece en el ejemplo 1.3 ilustra de manera adecuada los tipos de preguntas de ingeniería que se plantean y se responden aplicando los métodos estadísticos que se utilizan en un diseño experimental y se presentan en este texto. Tales preguntas son las siguientes:

i. ¿Cuál es la naturaleza del efecto de la humedad relativa sobre la corrosión de la aleación de aluminio dentro del rango de humedad relativa en este experimento?

ii. ¿El recubrimiento químico contra la corrosión reduce los niveles de corrosión y existe alguna manera de cuantificar el efecto?

iii. ¿Hay alguna **interacción** entre el tipo de recubrimiento y la humedad relativa que influya en la corrosión de la aleación? Si es así, ¿cómo se podría interpretar?

¿Qué es interacción?

La importancia de las preguntas **i.** y **ii.** debería quedar clara para el lector, ya que ambas tienen que ver con aspectos importantes tanto para los productores como para los usuarios de la aleación. ¿Pero qué sucede con la pregunta **iii.**? El concepto de *interacción* se estudiará con detalle en los capítulos 14 y 15. Considere la gráfica de la figura 1.3, la cual ejemplifica la detección de la interacción entre dos **factores** en un diseño experimental simple. Observe que las líneas que conectan las medias de la muestra no son paralelas. El **paralelismo** habría indicado que el efecto (visto como un resultado de la pendiente de las líneas) de la humedad relativa es igual, es decir, negativo, tanto en la condición sin recubrimiento como en la condición con recubrimiento químico contra la corrosión. Recuerde que la pendiente negativa implica que la corrosión se vuelve más pronunciada a medida que aumenta la humedad. La ausencia de paralelismo implica una interacción entre el tipo de recubrimiento y la humedad relativa. La línea casi "horizontal" para el recubrimiento contra la corrosión, opuesta a la pendiente más pronunciada para la condición sin recubrimiento, sugiere que *el recubrimiento químico contra la corrosión no sólo es benéfico (observe el desplazamiento entre las líneas), sino que la presencia del recubrimiento revela que el efecto de la humedad es despreciable*. Salta a la vista que todas estas cuestiones son muy importantes para el efecto de los dos factores individuales y para la interpretación de la interacción, si está presente.

Los modelos estadísticos son muy útiles para responder preguntas como las descritas en i, ii y iii, en donde los datos provienen de un diseño experimental. Sin embargo, no siempre se cuenta con el tiempo o los recursos que permiten usar un diseño experimental. Por ejemplo, hay muchos casos en los que las condiciones de interés para el científico o el ingeniero simplemente no se pueden implementar *debido a que es imposible controlar los factores importantes*. En el ejemplo 1.3 la humedad relativa y el tipo de recubrimiento (o la ausencia de éste) son bastante fáciles de controlar. Desde luego, se trata del rasgo distintivo de un diseño experimental. En muchos campos los factores a estudiar no pueden ser controlados por diversas razones. Un control riguroso como el del ejemplo 1.3 permite al analista confiar en que las diferencias encontradas (como en los niveles de

www.FreeLibros.me

corrosión) se deben a los factores que se pueden controlar. Considere el ejercicio 1.6 de la página 13 como otro ejemplo. En este caso suponga que se eligen 24 especímenes de caucho de silicio y que se asignan 12 a cada uno de los niveles de temperatura de vulcanizado. Las temperaturas se controlan cuidadosamente, por lo que éste es un ejemplo de diseño experimental con **un solo factor**, que es la temperatura de vulcanizado. Se podría suponer que las diferencias encontradas en la media de la resistencia a la tensión son atribuibles a las diferentes temperaturas de vulcanizado.

¿Qué sucede si no se controlan los factores?

Suponga que los factores no se controlan y que *no hay asignación aleatoria* a los tratamientos específicos para las unidades experimentales, y que se necesita obtener información a partir de un conjunto de datos. Como ejemplo considere un estudio donde el interés se centra en la relación entre los niveles de colesterol sanguíneo y la cantidad de sodio medida en la sangre. Durante cierto periodo se revisó el colesterol sanguíneo y el nivel de sodio de un grupo de individuos. En efecto, es posible obtener alguna información útil de tal conjunto de datos. Sin embargo, debería quedar claro que no es posible hacer un control estricto de los niveles de sodio. De manera ideal, los sujetos deberían dividirse aleatoriamente en dos grupos, donde uno fuera el asignado a un nivel alto específico de sodio en la sangre, y el otro a un nivel bajo específico de sodio en la sangre, pero es obvio que esto no es posible. Evidentemente los cambios en los niveles de colesterol se deben a cambios en uno o diversos factores que no se controlaron. Este tipo de estudio, sin control de factores, se denomina **estudio observacional**, el cual la mayoría de las veces implica una situación en que los sujetos se observan a través del tiempo.

Los estudios biológicos y biomédicos a menudo tienen que ser observacionales. Sin embargo, este tipo de estudios no se restringen a dichas áreas. Por ejemplo, considere un estudio diseñado para determinar la influencia de la temperatura ambiental sobre la energía eléctrica que consumen las instalaciones de una planta química. Es evidente que los niveles de la temperatura ambiental no se pueden controlar, por lo tanto, la única manera en que se puede supervisar la estructura de los datos es a partir de los datos de la planta a través del tiempo.

Es importante destacar que una diferencia básica entre un experimento bien diseñado y un estudio observacional es la dificultad para determinar la causa y el efecto verdaderos con este último. Asimismo, las diferencias encontradas en la reacción fundamental (por ejemplo, niveles de corrosión, colesterol sanguíneo, consumo de energía eléctrica en una planta) podrían deberse a otros factores subyacentes que no se controlaron. De manera ideal, en un diseño experimental los *factores perturbadores* serían compensados mediante el proceso de aleatoriedad. En realidad, los cambios en los niveles de colesterol sanguíneo podrían deberse a la ingestión de grasa, a la realización de actividad física, etc. El consumo de energía eléctrica podría estar afectado por la cantidad de bienes producidos o incluso por la pureza de éstos.

Otra desventaja de los estudios observacionales, que a menudo se ignora cuando éstos se comparan con experimentos cuidadosamente diseñados, es que, a diferencia de estos últimos, los observacionales están a merced de circunstancias no controladas, naturales, ambientales o de otros tipos, que repercuten en los niveles de los factores de interés. Por ejemplo, en el estudio biomédico acerca de la influencia de los niveles de sodio en la sangre sobre el colesterol sanguíneo es posible que haya, de hecho, una influencia significativa, pero el conjunto de datos específico usado no involucró la suficiente variación observada en los niveles de sodio debido a la naturaleza de los sujetos elegidos. Desde luego, en un diseño experimental el analista elige y controla los niveles de los factores.

www.FreeLibros.me

Un tercer tipo de estudio estadístico que podría ser muy útil, pero que tiene notables desventajas cuando se le compara con un diseño experimental, es el **estudio retrospectivo**. Esta clase de estudio emplea estrictamente **datos históricos**, que se obtienen durante un periodo específico. Una ventaja evidente de los datos retrospectivos es el bajo costo de la recopilación de datos. Sin embargo, como se podría esperar, también tiene desventajas claras:

i. La validez y la confiabilidad de los datos históricos a menudo son cuestionables.

ii. Si el tiempo es un aspecto relevante en la estructura de los datos, podría haber datos faltantes.

iii. Podrían existir errores en la recopilación de los datos que no se conocen.

iv. De nuevo, como en el caso de los datos observacionales, no hay control en los rangos de las variables a medir (es decir, en los factores a estudiar). De hecho, las variaciones que se encuentran en los datos históricos a menudo no son significativas para estudios actuales.

En la sección 1.6 se puso cierto énfasis en los modelos de las relaciones entre variables. Presentamos el concepto de análisis de regresión, el cual se estudia en los capítulos 11 y 12, y se considera como una forma del análisis de datos para los diseños experimentales que se examinarán en los capítulos 14 y 15. En la sección 1.6 se utilizó a modo de ejemplo un modelo que relaciona la media poblacional de la resistencia a la tensión de la tela con los porcentajes de algodón, en el cual 20 especímenes de tela representaban las unidades experimentales. En este caso, los datos provienen de un diseño experimental simple, en el que los porcentajes de algodón individuales fueron seleccionados por el científico.

Con frecuencia, tanto los datos observacionales como los retrospectivos se utilizan para observar las relaciones entre variables a través de los procedimientos de construcción de modelos que se estudiarán en los capítulos 11 y 12. Aunque las ventajas de los diseños experimentales se pueden aplicar cuando la finalidad es la construcción de un modelo estadístico, hay muchas áreas en las que no es posible diseñar experimentos, de manera que *habrá que utilizar los datos históricos u observacionales*. Aquí nos referimos al conjunto de datos históricos que se incluye en el ejercicio 12.5 de la página 450. El objetivo es construir un modelo que dé como resultado una ecuación o relación que vincule el consumo mensual de energía eléctrica con la temperatura ambiental promedio, x_1, el número de días en el mes, x_2, la pureza promedio del producto, x_3 y las toneladas de bienes producidos, x_4. Se trata de los datos históricos del año anterior.

Ejercicios

1.13 Un fabricante de componentes electrónicos se interesa en determinar el tiempo de vida de cierto tipo de batería. Una muestra, en horas de vida, es como la siguiente:

123, 116, 122, 110, 175, 126, 125, 111, 118, 117.

a) Calcule la media y la mediana de la muestra.
b) ¿Qué característica en este conjunto de datos es la responsable de la diferencia sustancial entre ambas?

1.14 Un fabricante de neumáticos quiere determinar el diámetro interior de un neumático de cierto grado de calidad. Idealmente el diámetro sería de 570 mm. Los datos son los siguientes:

572, 572, 573, 568, 569, 575, 565, 570.

a) Calcule la media y la mediana de la muestra.
b) Obtenga la varianza, la desviación estándar y el rango de la muestra.
c) Con base en los estadísticos calculados en los incisos *a*) y *b*), ¿qué comentaría acerca de la calidad de los neumáticos?

1.15 Cinco lanzamientos independientes de una moneda tienen como resultado *cinco caras.* Resulta que si la moneda es legal, la probabilidad de este resultado es $(1/2)^5 = 0.03125$. ¿Proporciona esto evidencia sólida

www.FreeLibros.me

de que la moneda no es legal? Comente y utilice el concepto de valor-*P* que se analizó en la sección 1.1.

1.16 Muestre que las n piezas de información en $\sum_{i=1}^{n}(x_i - \bar{x})^2$ no son independientes; es decir, demuestre que

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

1.17 Se realiza un estudio acerca de los efectos del tabaquismo sobre los patrones de sueño. La medición que se observa es el tiempo, en minutos, que toma quedar dormido. Se obtienen los siguientes datos:

Fumadores:	69.3	56.0	22.1	47.6
	53.2	48.1	52.7	34.4
	60.2	43.8	23.2	13.8
No fumadores:	28.6	25.1	26.4	34.9
	29.8	28.4	38.5	30.2
	30.6	31.8	41.6	21.1
	36.0	37.9	13.9	

a) Calcule la media de la muestra para cada grupo.
b) Calcule la desviación estándar de la muestra para cada grupo.
c) Elabore una gráfica de puntos de los conjuntos de datos A y B en la misma línea.
d) Comente qué clase de efecto parece tener el hecho de fumar sobre el tiempo que se requiere para quedarse dormido.

1.18 Las siguientes puntuaciones representan la calificación en el examen final para un curso de estadística elemental:

23	60	79	32	57	74	52	70	82
36	80	77	81	95	41	65	92	85
55	76	52	10	64	75	78	25	80
98	81	67	41	71	83	54	64	72
88	62	74	43	60	78	89	76	84
48	84	90	15	79	34	67	17	82
69	74	63	80	85	61			

a) Elabore un diagrama de tallo y hojas para las calificaciones del examen, donde los tallos sean 1, 2, 3,…, 9.
b) Elabore un histograma de frecuencias relativas, trace un estimado de la gráfica de la distribución y analice la asimetría de la distribución.
c) Calcule la media, la mediana y la desviación estándar de la muestra.

1.19 Los siguientes datos representan la duración de vida, en años, medida al entero más cercano, de 30 bombas de combustible similares.

2.0	3.0	0.3	3.3	1.3	0.4
0.2	6.0	5.5	6.5	0.2	2.3
1.5	4.0	5.9	1.8	4.7	0.7
4.5	0.3	1.5	0.5	2.5	5.0
1.0	6.0	5.6	6.0	1.2	0.2

a) Construya un diagrama de tallo y hojas para la vida, en años, de las bombas de combustible, utilizando el dígito a la izquierda del punto decimal como el tallo para cada observación.
b) Determine una distribución de frecuencias relativas.
c) Calcule la media, el rango y la desviación estándar de la muestra.

1.20 Los siguientes datos representan la duración de la vida, en segundos, de 50 moscas de la fruta que se someten a un nuevo aerosol en un experimento de laboratorio controlado.

17	20	10	9	23	13	12	19	18	24
12	14	6	9	13	6	7	10	13	7
16	18	8	3	3	32	9	7	10	11
13	7	18	7	10	4	27	19	16	8
7	10	5	14	15	10	9	6	7	15

a) Elabore un diagrama doble de tallo y hojas para el periodo de vida de las moscas de la fruta usando los tallos 0★, 0•, 1★, 1•, 2★, 2• y 3★ de manera que los tallos codificados con los símbolos ★ y • se asocien, respectivamente, con las hojas 0 a 4 y 5 a 9.
b) Determine una distribución de frecuencias relativas.
c) Construya un histograma de frecuencias relativas.
d) Calcule la mediana.

1.21 La duración de fallas eléctricas, en minutos, se presenta en la siguiente tabla.

22	18	135	15	90	78	69	98	102
83	55	28	121	120	13	22	124	112
70	66	74	89	103	24	21	112	21
40	98	87	132	115	21	28	43	37
50	96	118	158	74	78	83	93	95

a) Calcule la media y la mediana muestrales de las duraciones de la falla eléctrica.
b) Calcule la desviación estándar de las duraciones de la falla eléctrica.

1.22 Los siguientes datos son las mediciones del diámetro de 36 cabezas de remache en centésimos de una pulgada.

6.72	6.77	6.82	6.70	6.78	6.70	6.62	6.75
6.66	6.66	6.64	6.76	6.73	6.80	6.72	6.76
6.76	6.68	6.66	6.62	6.72	6.76	6.70	6.78
6.76	6.67	6.70	6.72	6.74	6.81	6.79	6.78
6.66	6.76	6.76	6.72				

a) Calcule la media y la desviación estándar de la muestra.
b) Construya un histograma de frecuencias relativas para los datos.

www.FreeLibros.me

c) Comente si existe o no una indicación clara de que la muestra proviene de una población que tiene una distribución en forma de campana.

1.23 En 20 automóviles elegidos aleatoriamente, se tomaron las emisiones de hidrocarburos en velocidad en vacío, en partes por millón (ppm), para modelos de 1980 y 1990.

Modelos 1980:

141 359 247 940 882 494 306 210 105 880
200 223 188 940 241 190 300 435 241 380

Modelos 1990:

140 160 20 20 223 60 20 95 360 70
220 400 217 58 235 380 200 175 85 65

a) Construya una gráfica de puntos como la de la figura 1.1.
b) Calcule la media de la muestra para los dos años y sobreponga las dos medias en las gráficas.
c) Comente sobre lo que indica la gráfica de puntos respecto de si cambiaron o no las emisiones poblacionales de 1980 a 1990. Utilice el concepto de variabilidad en sus comentarios.

1.24 Los siguientes son datos históricos de los sueldos del personal (dólares por alumno) en 30 escuelas seleccionadas de la región este de Estados Unidos a principios de la década de 1970.

3.79 2.99 2.77 2.91 3.10 1.84 2.52 3.22
2.45 2.14 2.67 2.52 2.71 2.75 3.57 3.85
3.36 2.05 2.89 2.83 3.13 2.44 2.10 3.71
3.14 3.54 2.37 2.68 3.51 3.37

a) Calcule la media y la desviación estándar de la muestra.
b) Utilice los datos para elaborar un histograma de frecuencias relativas.
c) Construya un diagrama de tallo y hojas con los datos.

1.25 El siguiente conjunto de datos se relaciona con el ejercicio 1.24 y representa el porcentaje de las familias que se ubican en el nivel superior de ingresos en las mismas escuelas individuales y con el mismo orden del ejercicio 1.24.

72.2 31.9 26.5 29.1 27.3 8.6 22.3 26.5
20.4 12.8 25.1 19.2 24.1 58.2 68.1 89.2
55.1 9.4 14.5 13.9 20.7 17.9 8.5 55.4
38.1 54.2 21.5 26.2 59.1 43.3

a) Calcule la media de la muestra.
b) Calcule la mediana de la muestra.
c) Construya un histograma de frecuencias relativas con los datos.
d) Determine la media recortada al 10%. Compárela con los resultados de los incisos *a*) y *b*) y exprese su comentario.

1.26 Suponga que le interesa emplear los conjuntos de datos de los ejercicios 1.24 y 1.25 para derivar un modelo que prediga los salarios del personal como una función del porcentaje de familias en un nivel alto de ingresos para los sistemas escolares actuales. Comente sobre cualquier desventaja de llevar a cabo este tipo de análisis.

1.27 Se realizó un estudio para determinar la influencia del desgaste, y, de un cojinete como una función de la carga, x, sobre el cojinete. Para este estudio se utilizó un diseño experimental con tres niveles de carga: 700 lb, 1000 lb y 1300 lb. En cada nivel se utilizaron cuatro especímenes y las medias muestrales fueron 210, 325 y 375, respectivamente.

a) Grafique el promedio de desgaste contra la carga.
b) A partir de la gráfica del inciso *a*), ¿consideraría que hay una relación entre desgaste y carga?
c) Suponga que tenemos los siguientes valores individuales de desgaste para cada uno de los cuatro especímenes en los respectivos niveles de carga. (Vea los datos que siguen). Grafique los resultados de desgaste para todos los especímenes contra los tres valores de carga.
d) A partir de la gráfica del inciso *c*), ¿consideraría que hay una relación clara? Si su respuesta difiere de la del inciso *b*), explique por qué.

	x		
	700	1000	1300
y_1	145	250	150
y_2	105	195	180
y_3	260	375	420
y_4	330	480	750
	$\bar{y}_1 = 210$	$\bar{y}_2 = 325$	$\bar{y}_3 = 375$

1.28 En Estados Unidos y otros países muchas compañías de manufactura utilizan partes moldeadas como componentes de un proceso. La contracción a menudo es un problema importante. Por consiguiente, un dado de metal moldeado para una parte se construye más grande que el tamaño nominal con el fin de permitir su contracción. En un estudio de moldeado por inyección se descubrió que en la contracción influyen múltiples factores, entre los cuales están la velocidad de la inyección en pies/segundo y la temperatura de moldeado en °C. Los dos conjuntos de datos siguientes muestran los resultados del diseño experimental, en donde la velocidad de inyección se mantuvo a dos niveles (bajo y alto) y la temperatura de moldeado se mantuvo constante en un nivel bajo. La contracción se midió en cm $\times 10^4$.

Los valores de contracción a una velocidad de inyección baja fueron:

72.68 72.62 72.58 72.48 73.07
72.55 72.42 72.84 72.58 72.92

Los valores de contracción a una velocidad de inyección alta fueron:

71.62 71.68 71.74 71.48 71.55
71.52 71.71 71.56 71.70 71.50

www.FreeLibros.me

a) Construya una gráfica de puntos para ambos conjuntos de datos en la misma gráfica. Sobre ésta indique ambas medias de la contracción, tanto para la velocidad de inyección baja como para la velocidad de inyección alta.
b) Con base en los resultados de la gráfica del inciso *a*), y considerando la ubicación de las dos medias y su sentido de variabilidad, ¿cuál es su conclusión respecto del efecto de la velocidad de inyección sobre la contracción a una temperatura de moldeado baja?

1.29 Utilice los datos del ejercicio 1.24 para elaborar una gráfica de caja.

1.30 A continuación se presentan los tiempos de vida, en horas, de 50 lámparas incandescentes, con esmerilado interno, de 40 watts y 110 voltios, los cuales se tomaron de pruebas forzadas de vida:

919	1196	785	1126	936	918
1156	920	948	1067	1092	1162
1170	929	950	905	972	1035
1045	855	1195	1195	1340	1122
938	970	1237	956	1102	1157
978	832	1009	1157	1151	1009
765	958	902	1022	1333	811
1217	1085	896	958	1311	1037
702	923				

Elabore una gráfica de puntos para estos datos.

1.31 Considere la situación del ejercicio 1.28, pero ahora utilice el siguiente conjunto de datos, en el cual la contracción se mide de nuevo a una velocidad de inyección baja y a una velocidad de inyección alta. Sin embargo, esta vez la temperatura de moldeado se aumenta a un nivel "alto" y se mantiene constante.

Los valores de la contracción a una velocidad de inyección baja fueron:

76.20	76.09	75.98	76.15	76.17
75.94	76.12	76.18	76.25	75.82

Los valores de la contracción a una velocidad de inyección alta fueron:

93.25	93.19	92.87	93.29	93.37
92.98	93.47	93.75	93.89	91.62

a) Igual que en el ejercicio 1.28, elabore una gráfica de puntos con ambos conjuntos de datos en la misma gráfica e identifique las dos medias (es decir, la contracción media para la velocidad de inyección baja y para la velocidad de inyección alta).
b) Igual que en el ejercicio 1.28, comente sobre la influencia de la velocidad de inyección en la contracción para la temperatura de moldeado alta. Tome en cuenta la posición de las dos medias y la variabilidad de cada media.
c) Compare su conclusión en el inciso *b*) actual con la del inciso *b*) del ejercicio 1.28, en el cual la temperatura de moldeado se mantuvo a un nivel bajo. ¿Diría que hay interacción entre la velocidad de inyección y la temperatura de moldeado? Explique su respuesta.

1.32 Utilice los resultados de los ejercicios 1.28 y 1.31 para crear una gráfica que ilustre la interacción evidente entre los datos. Use como guía la gráfica de la figura 1.3 del ejemplo 1.3. ¿El tipo de información que se encontró en los ejercicios 1.28 y 1.31 se habría encontrado en un estudio observacional en el que el analista no hubiera tenido control sobre la velocidad de inyección ni sobre la temperatura de moldeado? Explique su respuesta.

1.33 Proyecto de grupo: Registre el tamaño de calzado que usa cada estudiante de su grupo. Utilice las medias y las varianzas muestrales, así como los tipos de gráficas que se estudiaron en este capítulo, para resumir cualquier característica que revele una diferencia entre las distribuciones del tamaño del calzado de hombres y mujeres. Haga lo mismo con la estatura de cada estudiante de su grupo.

www.FreeLibros.me

www.FreeLibros.me

Capítulo 2

Probabilidad

2.1 Espacio muestral

En el estudio de la estadística tratamos básicamente con la presentación e interpretación de **resultados fortuitos** que ocurren en un estudio planeado o en una investigación científica. Por ejemplo, en Estados Unidos, y con la finalidad de justificar la instalación de un semáforo, se podría registrar el número de accidentes que ocurren mensualmente en la intersección de Driftwood Lane y Royal Oak Drive; en una fábrica se podrían clasificar los artículos que salen de la línea de ensamble como "defectuosos" o "no defectuosos"; en una reacción química se podría revisar el volumen de gas que se libera cuando se varía la concentración de un ácido. Por ello, quienes se dedican a la estadística a menudo manejan datos numéricos que representan conteos o mediciones, o **datos categóricos** que se podrían clasificar de acuerdo con algún criterio.

En este capítulo, al referirnos a cualquier registro de información, ya sea numérico o categórico, utilizaremos el término **observación**. Por consiguiente, los números 2, 0, 1 y 2, que representan el número de accidentes que ocurrieron cada mes, de enero a abril, durante el año pasado en la intersección de Driftwood Lane y Royal Oak Drive, constituyen un conjunto de observaciones. Lo mismo ocurre con los datos categóricos *N, D, N, N* y *D*, que representan los artículos defectuosos o no defectuosos cuando se inspeccionan cinco artículos y se registran como observaciones.

Los estadísticos utilizan la palabra **experimento** para describir cualquier proceso que genere un conjunto de datos. Un ejemplo simple de experimento estadístico es el lanzamiento de una moneda al aire. En tal experimento sólo hay dos resultados posibles: cara o cruz. Otro experimento podría ser el lanzamiento de un misil y la observación de la velocidad a la que se desplaza en tiempos específicos. Las opiniones de los votantes respecto de un nuevo impuesto sobre las ventas también se pueden considerar como observaciones de un experimento. En estadística nos interesan, en particular, las observaciones que se obtienen al repetir varias veces un experimento. En la mayoría de los casos los resultados dependerán del azar, por lo tanto, no se pueden predecir con certeza. Si un químico realizara un análisis varias veces en las mismas condiciones, obtendría diferentes medidas, las cuales indicarían un elemento de probabilidad en el procedimiento experimental. Aun cuando lancemos una moneda al aire repetidas veces, no podemos tener la certeza de que en un lanzamiento determinado obtendremos cara como resultado. Sin embargo, conocemos el conjunto completo de posibilidades para cada lanzamiento.

Dado lo expuesto en la sección 1.7, en la que se revisaron tres tipos de estudios estadísticos y se dieron varios ejemplos de cada uno, ya deberíamos estar familiarizados con el alcance del término experimento. En cada uno de los tres casos, *diseños experimentales*, *estudios observacionales* y *estudios retrospectivos*, el resultado final fue un conjunto

www.FreeLibros.me

de datos que, por supuesto, está sujeto a la **incertidumbre**. Aunque sólo uno de ellos tiene la palabra *experimento* en su descripción, el proceso de generar los datos o el proceso de observarlos forma parte de un experimento. El estudio de la corrosión expuesto en la sección 1.2 ciertamente implica un experimento en el que los datos son representados por las mediciones de la corrosión. El ejemplo de la sección 1.7, en el que se observó el colesterol y el sodio en la sangre de un conjunto de individuos, representó un estudio observacional (como lo opuesto a un *diseño* experimental) en el que el proceso incluso generó datos y un resultado sujeto a la incertidumbre; por lo tanto, se trata de un experimento. Un tercer ejemplo en la sección 1.7 consistió en un estudio retrospectivo, en el cual se observaron datos históricos sobre el consumo de energía eléctrica por mes y el promedio mensual de la temperatura ambiental. Aun cuando los datos pueden haber estado archivados durante décadas, el proceso se seguirá considerando un experimento.

Definición 2.1: Al conjunto de todos los resultados posibles de un experimento estadístico se le llama **espacio muestral** y se representa con el símbolo S.

A cada resultado en un espacio muestral se le llama **elemento** o **miembro** del espacio muestral, o simplemente **punto muestral**. Si el espacio muestral tiene un número finito de elementos, podemos *listar* los miembros separados por comas y encerrarlos entre llaves. Por consiguiente, el espacio muestral S, de los resultados posibles cuando se lanza una moneda al aire, se puede escribir como

$$S = \{H, T\},$$

en donde H y T corresponden a "caras" y "cruces", respectivamente.

Ejemplo 2.1: Considere el experimento de lanzar un dado. Si nos interesara el número que aparece en la cara superior, el espacio muestral sería

$$S_1 = \{1, 2, 3, 4, 5, 6\}$$

Si sólo estuviéramos interesados en si el número es par o impar, el espacio muestral sería simplemente

$$S_2 = \{\text{par, impar}\}$$

El ejemplo 2.1 ilustra el hecho de que se puede usar más de un espacio muestral para describir los resultados de un experimento. En este caso, S_1 brinda más información que S_2. Si sabemos cuál elemento ocurre en S_1, podremos indicar cuál resultado tiene lugar en S_2; no obstante, saber lo que pasa en S_2 no ayuda mucho a determinar qué elemento ocurre en S_1. En general, lo deseable sería utilizar un espacio muestral que proporcione la mayor información acerca de los resultados del experimento. En algunos experimentos es útil listar los elementos del espacio muestral de forma sistemática utilizando un **diagrama de árbol**.

Ejemplo 2.2: Un experimento consiste en lanzar una moneda y después lanzarla una segunda vez si sale cara. Si en el primer lanzamiento sale cruz, entonces se lanza un dado una vez. Para listar los elementos del espacio muestral que proporciona la mayor información construimos el diagrama de árbol de la figura 2.1. Las diversas trayectorias a lo largo de las ramas del árbol dan los distintos puntos muestrales. Si empezamos con la rama superior izquierda y nos movemos a la derecha a lo largo de la primera trayectoria, obtenemos el punto muestral *HH,* que indica la posibilidad de que ocurran caras en dos lanzamientos sucesivos de la moneda. De igual manera, el punto muestral *T*3 indica la posibilidad de que la moneda muestre una cruz seguida por un 3 en el lanzamiento del dado. Al seguir todas las trayectorias, vemos que el espacio muestral es

www.FreeLibros.me

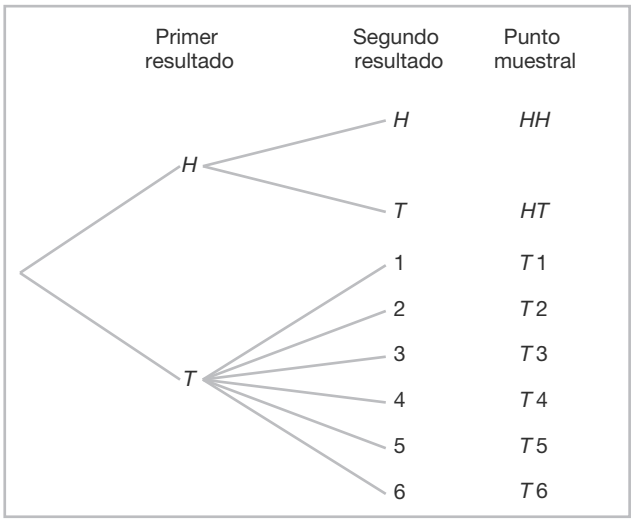


Figura 2.1: Diagrama de árbol para el ejemplo 2.2.

$$S = \{HH, HT, T1, T2, T3, T4, T5, T6\}.$$

Muchos de los conceptos de este capítulo se ilustran mejor con ejemplos que involucran el uso de dados y cartas. Es particularmente importante utilizar estas aplicaciones al comenzar el proceso de aprendizaje, ya que facilitan el flujo de esos conceptos nuevos en ejemplos científicos y de ingeniería como el siguiente.

Ejemplo 2.3: Suponga que se seleccionan, de forma aleatoria, tres artículos de un proceso de fabricación. Cada artículo se inspecciona y se clasifica como defectuoso, *D,* o no defectuoso, *N*. Para listar los elementos del espacio muestral que brinde la mayor información, construimos el diagrama de árbol de la figura 2.2, de manera que las diversas trayectorias a lo largo de las ramas del árbol dan los distintos puntos muestrales. Al comenzar con la primera trayectoria, obtenemos el punto muestral *DDD,* que indica la posibilidad de que los tres artículos inspeccionados estén defectuosos. Conforme continuamos a lo largo de las demás trayectorias, vemos que el espacio muestral es

$$S = \{DDD, DDN, DND, DNN, NDD, NDN, NND, NNN\}.$$

Los espacios muestrales con un número grande o infinito de puntos muestrales se describen mejor mediante un **enunciado** o **método de la regla**. Por ejemplo, si el conjunto de resultados posibles de un experimento fuera el conjunto de ciudades en el mundo con una población de más de un millón de habitantes, nuestro espacio muestral se escribiría como

$$S = \{x \mid x \text{ es una ciudad con una población de más de un millón de habitantes}\},$$

que se lee "*S* es el conjunto de todas las *x,* tales que *x* es una ciudad con una población de más de un millón de habitantes". La barra vertical se lee como "tal que". De manera similar, si *S* es el conjunto de todos los puntos (x, y) sobre los límites o el interior de un círculo de radio 2 con centro en el origen, escribimos la **regla**

$$S = \{(x, y) \mid x^2 + y^2 \leq 4\}.$$

www.FreeLibros.me